
GAZİ ÜNİVERSİTESİ
Gazi
1926

EARLY DETECTION OF CEREBRAL ISCHEMIA AND HEMORRHAGE USING ARTIFICIAL NEURAL NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
GAZİ UNIVERSITY

BY

Anıl AKYEL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL-ELECTRONICS ENGINEERING

JUNE 2021

ETHICAL STATEMENT

I hereby declare that in this thesis study I prepared in accordance with thesis writing rules of Gazi University Graduate School of Natural and Applied Sciences;

- All data, information and documents presented in this thesis have been obtained within the scope of academic rules and ethical conduct,
- All information, documents, assessments and results have been presented in accordance with scientific ethical conduct and moral rules,
- All material used in this thesis that are not original to this work have been fully cited and referenced,
- No change has been made in the data used,
- The work presented in this thesis is original,

or else, I admit all loss of rights to be incurred against me.

Anıl AKYEL

25/06/2021

EARLY DETECTION OF CEREBRAL ISCHEMIA AND HEMORRHAGE USING ARTIFICIAL NEURAL NETWORKS

(Ph. D. Thesis)

Anıl AKYEL

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2021

ABSTRACT

Stroke, which has two main types as cerebral ischemia and hemorrhage, is a serious health condition that is among the leading causes of death and permanent disability worldwide. Despite this fact, a generally valid, effective treatment method is yet to be found in the struggle against stroke. This situation has led to the concentration of studies on the prediction of possible attacks in order to minimize the impact of stroke events. However, the acute nature of stroke events and the high number of factors affecting the risk of stroke, some of which are still unknown, make it difficult to establish an accurate risk estimation method. A novel, 3000-patient dataset is formed and used in training and testing of the proposed model within the context of study, using the data obtained from Gazi University Hospital retrospectively. The model proposed in the study consists of 2 main stages. In the first stage, the dataset input to the model is clustered using fuzzy c-means algorithm and the clustering results obtained are used to modify the dataset. In the second stage, this modified dataset is taken as input and the risks of the patients in the set for 4 different stroke types are estimated with ensemble learning based classifiers that consist of 4 different base models for each stroke type. With a novel approach, the weights of the base models used in this stage are determined dynamically according to their performances in validation dataset, separately for each cluster. By this way, not only a high model performance is obtained, but also the operating time and computational load of the model is reduced considerably. Furthermore, the clustering process and the usage of different weights in the resulting clusters provided the model with resistance against possible biases in the datasets. The accuracy of the model in predicting general stroke, ischemic stroke, hemorrhagic stroke and transient ischemic attack events are obtained as 99,89%; 99,90%; 99,81% and 99,87%; respectively. When these accuracy values are considered, it is observed that the model proposed in the study shows almost perfect performance. In this context, this model, with its ability to perform risk estimation accurately for 4 different types of stroke by considering a large number of factors can aid the medical experts by functioning as a decision support system.

Science Code	:	90542
Key Words	:	Artificial intelligence, stroke, fuzzy clustering, ensemble learning
Page Number	:	160
Supervisor	:	Prof. Dr. Fırat HARDALAÇ
Co-Supervisor	:	Prof. Dr. Pınar AKDEMİR ÖZIŞIK

YAPAY SİNİR AĞLARI KULLANILARAK SEREBRAL İSKEMİ VE HEMORAJİ OLAYLARININ ÖNCEDEN TAHMİN EDİLMESİ

(Doktora Tezi)

Anıl AKYEL

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Haziran 2021

ÖZET

İnme, serebral iskemi ve hemoraji şeklinde iki ana alt türü bulunan ve dünya çapında ölüm ve kalıcı sakatlıkların önde gelen sebeplerinden biri olan ciddi bir sağlık sorunudur. Buna karşın, inme ile mücadelede bugüne kadar genel geçer bir etkin tedavi yöntemi bulunamamıştır. Bu durum da inme olaylarının etkisini minimize etmek için olası atakların önceden tahmin edilmesi konusunda çalışmalara yoğunlaşılmasına yol açmıştır. Ancak, inme olaylarının akut doğası ve bireylerin inme geçirme risklerini etkileyen faktörlerin, bazıları hâlâ bilinmemekle birlikte, sayıca çok olması başarılı bir risk tahmin yöntemi oluşturmayı oldukça zor kılmaktadır. Bu çalışmada, bireylerin inme geçirme risklerini hesaplayarak yüksek risk altındaki kişilerin belirlenmesi ve meydana gelebilecek ataklara yönelik gerekli önlemlerin alınması konusunda bir karar-destek sistemi olarak işlev görecek bir model önerilmiştir. Önerilen model, Matlab yazılımı kullanılarak hazırlanmış olup kullanım kolaylığı ve anlaşılabilirlik için bir grafiksel kullanıcı arayüzü ile tamamlanmıştır. Çalışma kapsamında Gazi Üniversitesi Hastanesi'nden retrospektif olarak alınan 3000 hasta verisi kullanılarak özgün bir veri seti hazırlanmış olup önerilen modelin eğitimi ve testinde bu veri seti kullanılmıştır. Çalışmada önerilen model 2 ana aşamadan oluşmaktadır. İlk aşamada modele verilen veri seti bulanık c-means yöntemi ile kümelenerek elde edilen kümeleme sonuçları veri setini modifiye etmekte kullanılmıştır. İkinci aşamada ise modifiye edilen veri seti girdi olarak alınıp, setteki hastaların 4 farklı inme türü için riskleri her biri 4 farklı temel modelden oluşan topluluk öğrenme temelli sınıflandırıcılar ile tahmin edilmiştir. Burada kullanılan temel modellerin ağırlıkları özgün bir yaklaşımla, her küme için ayrı ayrı olmak üzere, temel modellerin doğrulama veri setindeki başarılarına göre dinamik olarak belirlenmiştir. Bu sayede hem yüksek bir model performansı elde edilmiş olup hem de modelin çalışma süresi ve işlem yükü kayda değer derecede azaltılmıştır. Ayrıca, uygulanan kümeleme işlemi ve bunun neticesinde elde edilen kümelerde farklı ağırlıkların kullanılması modele kullanılan veri setlerindeki olası yanlılıklara karşı direnç kazandırmıştır. Modelin genel inme, iskemik inme, hemorajik inme ve geçici iskemik atak olaylarını tahmin etmedeki doğruluk oranları sırasıyla %99,89; %99,90; %99,81 ve %99,87 olarak elde edilmiştir. Bu doğruluk oranları göz önünde bulundurulduğunda çalışmada önerilen modelin mükemmele yakın performansla çalıştığı gözlenmektedir. Bu bağlamda, çok sayıda faktörü hesaba katarak 4 farklı inme türü için yüksek doğruluk oranı ile risk tahmini yapabilen bu model, tıbbî uzmanlara bir karar destek sistemi vazifesi görerek yardımcı olabilecektir.

Bilim Kodu	:	90542
Anahtar Kelimeler	:	Yapay zeka, inme, bulanık kümeleme, topluluk öğrenme
Sayfa Adedi	:	160
Danışman	:	Prof. Dr. Fırat HARDALAÇ
İkinci Danışman	:	Prof. Dr. Pınar AKDEMİR ÖZIŞIK

ACKNOWLEDGMENTS

It is hard to describe how much I am grateful for the support and love from my dearest family, especially my mother with her endless and unconditional love, patience and care that made realization of this study possible.

I would like to express my gratitude to my thesis supervisor Prof. Dr. Fırat HARDALAÇ for his guidance throughout this study.

I would also like to thank my thesis monitoring committee members, Prof. Dr. Sabri KOÇER and Prof. Dr. Ferda Nur ALPASLAN for their guidance and advice during the conceptualization and design stages of this study.

Furthermore, I am grateful to Prof. Dr. Bijen NAZLIEL and Assoc. Prof. Dr. Hale BATUR ÇAĞLAYAN for their help and support during the data acquisition stage of this study. I am also thankful to Gazi University Hospital management for their assistance at this stage.

While they are unable to read, I would still like to mention my gratitude to my four legged sons, Bızdık and Sarman as they made my days bearable with their unconditional love and amusing antics during this long course.

Finally, I present my sincere gratitude to TÜBİTAK for the scholarship granted regarding my Ph. D. study.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

Table **Page**

Table **Page**

LIST OF FIGURES

Figure **Page**

Figure **Page**

SYMBOLS AND ABBREVIATIONS

The symbols and abbreviations used in this study are presented below with their explanations.

Symbols	Explanations
mmHg	millimeter of mercury
g	gram
mV	millivolt
mmol/L	millimoles per liter
ml/1,73min x m2	milliliter per 1,73 meter square per minute
mm^3	millimeter cube
kg/m^2	kilograms per meter square
kB	kilobytes
GHz	Gigahertz

Abbreviations	Explanations
AI	Artificial intelligence
AKI	Acute Kidney Injury
ALT	Alanine Aminotransferase
ANN	Artificial Neural Network
apTT	Activated Partial Thromboplastin Time
ASKH	Atherosclerotic heart disease
AST	Aspartate Transaminase
AUV	Autonomous Underwater Vehicle
AVM	Arteriovenous Malformation
BMI	Body Mass Index
BNP	B type Natriuretic Peptide
bpm	Beats Per Minute
BUN	Blood Urea Nitrogen
CAD-RADS	Coronary Artery Disease Reporting and Data System

Abbreviations	Explanations
CK	Creatine Kinase
CNN	Convolutional Neural Network
CPRD	Clinical Practice Research Datalink
CPU	Central Processing Unit
CRP	C-Reactive Protein
CT	Computer Tomography
CVD	Cardiovascular Disease
DBP	Diastolic Blood Pressure
DM	Diabetes Mellitus
EHR	Electronic Health Record
EM	Expectation Maximization
FC	Functional Connectivity
FCM	Fuzzy c-Means
GAN	Generative Adversarial Network
GDE	Graphical Design Environment
GGT	Gamma-Glutamyl Transferase
GUI	Graphical User Interface
HDL	High Density Lipoprotein
ICH	Intracerebral Hemorrhage
IHTRE	Ischemic, Hemorrhagic and TIA Risk Estimation
IST	International Stroke Trial
kNN	k-Nearest Neighbors
LDL	Low Density Lipoprotein
LSTM	Long Short-Term Memory
MCR	Margin Based Censored Regression
MCV	Mean Corpuscular Volume
MDL	Maximum Description Length
MDMA	3,4-Methylenedioxymethamphetamine
MI	Myocard Infarctus
MRI	Magnetic Resonance Imaging
Ms	Microsoft
MwA	Migraine with Auora

Abbreviations	Explanations
MwoA	Migraine without Auora
NHS	National Health Service
NIHSS	National Institute of Health Stroke Scale
OC	Oral Contraceptives
OCSP	Oxford Community Stroke Project classification
PCA	Posterior Cerebral Artery
PCOM	Posterior Communicating Artery
PDW	Platelet Distribution Width
ppd	Packs Per Day
PPM	Prosthesis–Patient Mismatch
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SAH	Subarachnoid Hemorrhage
SBP	Systolic Blood Pressure
SID	Synthesis of Integral Design
SVM	Support Vector Machine
TANN	Time-Aware Artificial Neural Network
TDI	Townsend Deprivation Index
TIA	Transient Ischemic Attack
TSH	Thyroid Stimulating Hormone
USA	United States of America
VLDL	Very Low Density Lipoprotein
VPC	Ventricular Premature Contractions
WSO	World Stroke Organization

1. INTRODUCTION

In this study an expert system based on artificial intelligence is proposed to perform stroke risk estimation on a given patient dataset. The expert system consists of two main modules, namely; clustering and risk estimation modules named after the tasks they perform.

Stroke Risk Estimation

Stroke is a medical condition that refers to the deterioration of blood flow in brain [1,2]. Due to its sudden and violent nature, stroke is considered as a hard-to-predict event in medical literature [3]. The fact that there have been relatively few biomarkers that might aid the medical professional in predicting the patients' risk of stroke [4] make the attacks more unpredictable, even with today's diagnostic tools. This unpredictable nature of stroke, combined with serious medical consequences on a considerable portion of patients worldwide [5,6] impose a considerable need for methods that could estimate the risk of stroke events in patients.

A number of different approaches for stroke risk estimation exist in this field. A commonly observed approach is to create a score chart in which a score is assigned to each factor or symptom considered [6]. These studies involve linear calculation of the scores based on whether a given symptom or factor is present for the patient in consideration. While some studies perform the risk calculation as a nonlinear combination of the scores of each factor, these studies are still based on predetermined scores, weights of which are usually chosen by the creators based on their medical experience. Usage of artificial intelligence in estimating the stroke risk of patients is a relatively new approach that has few examples in literature [7].

Aim of the Study

In this study, it is aimed to develop an expert system that can accurately and reliably estimate the risk of stroke in patients according to the patient data provided. The system is expected to not only estimate the general stroke risk, but also to estimate the type of expected stroke event by providing risk estimates for different types of stroke. This process involves the

estimation of risks of ischemic, hemorrhagic and transient ischemic (TIA) attacks, which are considered as different types of stroke events [8].

In order to effectively estimate the stroke risk in patients, the mechanism and factors of stroke are needed to be understood. To perform accurate estimations, the designed system must consider all factors that might have an effect on stroke risk of a given individual. Moreover, in order to achieve a high performance by means of both accuracy and reliability, a novel approach employing both supervised and unsupervised learning methods in two different stages of the algorithm is determined to be adopted. The final product is expected to allow the user flexibility in estimation of the risks of stroke types based on the results of the general stroke risk estimation process. Therefore, two different estimation modes are implemented in the final product allowing the user to trade some accuracy for speed whenever desired. Extensive tests are required to determine and optimize the performance of the system. For this reason, detailed analysis of each important parameter in proposed model is necessary to be carried out. The final product has a fully functional user interface, designed by bearing the fact that the final users would be mostly medical professionals. Therefore, an easy to understand and operate design is preferred over complicated, code-based operation.

Significance of the Study

Stroke is considered to have a large number of risk factors, the extent and mechanism of some are yet to be understood completely [9]. While certain factors such as diabetes, age or family history are known to affect the risk of stroke in individuals considerably, it is believed that two individuals with similar features might still have a large difference in their stroke risks [9] and such phenomenon cannot be detected by simply comparing the features of patients.

Stroke is the second most common cause of death worldwide, with 5.5 million deaths in 2016 [10]. It is estimated that around 20% of the patients lose their lives within the month following the stroke event [11]. This ratio increases to 50% for one year follow-up of stroke event. Moreover, even for those surviving a stroke event, the risk of becoming dependent (disabled) is reported to be around 67% [12]. Considering these facts, it is believed that a reliable system that could estimate the risk of stroke events in patients and warn the user

regarding any high-risk case would make it possible to take preventive precautions, even for patients with seemingly normal features. Since there is no known vaccine or drug to effectively prevent stroke events, taking preventive actions are considered as the most effective method in dealing with stroke. It is suggested that around 80% of stroke events are actually preventable by taking the necessary precautions [13] should the patients with high stroke risk are determined. Discriminating the patients with a high risk of stroke, as well determining the type of stroke that is more likely to occur would strengthen medical community's hand in their struggle with stroke, contributing to prevention of previously unexpected stroke events, possibly saving lives of many individuals. Such a contribution is believed to be invaluable as it would be in the direction of realizing one of the most honourable aims of science: benefiting humanity.

The large and incomplete list of possible stroke factors make it unlikely to achieve a high accuracy in estimating the stroke risks of patients with different sets of features for conventional stroke risk estimation charts. It is observed that stroke-related works involving artificial intelligence generally focus on detecting the stroke type and/or location after it has occurred, instead of estimating the risk of different stroke events. Few studies that employed machine learning algorithms to predict the stroke risk are noted in literature. Despite reportedly performing better than doctors in the experiments carried out, the performance reported in such studies were not very high [7, 14]. For this reason, implementing a reliable system employing a novel approach in artificial intelligence and machine learning that could provide the estimates of stroke risk in patients with a high accuracy while considering a large number of features is believed to be both an essential and novel contribution to the literature.

The raw data for the dataset created and used in this study is obtained from Gazi University Hospital in a retrospective manner following the approval of the ethics commission dated 09.12.2019 and is processed to a form that could be read by the developed system. This dataset is employed in developing, training and testing stages of the proposed model. It should be kept in mind that the dataset is based on diagnosis and examinations carried out by personnel of the hospital between years 2009-2019 and might have imperfect properties by means of values such as test results or patient anamnesis. Since it is impossible to detect or remove such imperfections which could possibly be present in any medical dataset, the dataset is accepted as it is.

Preliminary Definitions

In order to provide a better understanding of the study, some commonly used terms are explained in accordance with their usage throughout the related work.

Early Detection of Stroke

Early detection of stroke refers to detection of patients with a high risk of stroke. With the state of the art technology, the mechanism of stroke is considered as sudden and it is unlikely for a system to actually predict the time of a future stroke event. For this reason, early detection mainly refers to detecting the samples that the event is likely to occur at sometime in near future.

Stroke risk

Stroke risk refers to the likelihood of a stroke event taking place in an individual. It is commonly given in a decimal or percentile form. Same notion is valid for risks of stroke types such as ischemic, hemorrhagic or transient ischemic (TIA).

Ischemic stroke and TIA

Ischemic stroke refers to the event of cerebral ischemia. The stroke events originating from any material blocking the blood flow to brain tissue is regarded as ischemic stroke. TIA can be considered as the less severe subtype of ischemic stroke as the symptoms usually disappear within minutes. More information can be found in Medical Background section.

Hemorrhagic stroke

Hemorrhagic stroke refers to the event of cerebral hemorrhage. The stroke events originating from rupture of a cerebral blood vessel, resulting in pooling of blood in brain tissue is regarded as hemorrhagic stroke. More information can be found in Medical Background section.

Artificial neural network (ANN)

Artificial neural networks are computing systems that are designed by mimicking the neural systems of humans. Despite very simple structures such as a single layer perceptrons are also considered as ANNs, the system developed in this study could also be referred to as a complex, multi-stage ANN.

Thesis outline

In Chapter 2, the medical and engineering background necessary for this study is provided. This chapter also includes the literature review, presenting an overview of the past studies in related fields. Chapter 3 introduces the dataset formed and used in this study. In Chapter 4, the formation and operation of the proposed model is discussed in detail. Chapter 5 consists of the results of the analyzes performed in this study and their discussion. Finally, in Chapter 6, the thesis is concluded and ideas about future work is presented.

2. BACKGROUND AND LITERATURE REVIEW

In this study, an expert system employing artificial neural network approaches is implemented to be used in estimation of stroke risk in patients. While the method followed is essentially an engineering approach, medical background was necessary in many steps of the study such as understanding the factors of stroke, forming and processing the dataset, determining the multistage approach as well as interpreting the results. For this reason, a medical background is also provided alongside the engineering background. Later, a review of previous works in this field is presented.

2.1. Medical Background

Determination of stroke risks in patients is the main aim of this study. In order to do so, the fundamental information regarding stroke and its mechanism is needed to be understood. Furthermore, as the system is expected to estimate the risks of different types of stroke, the properties of these types as well as their similarities and differences should be understood. The consequences of stroke events are necessary to understand the importance of proposing such a system. Therefore, some important statistical data is also included in this section. Lastly, the most significant stroke risk factors are explained briefly to provide a better understanding of the features determined at the dataset formation stage.

2.1.1. Stroke

Stroke can be described as a medical condition that involves the deterioration of blood flow in brain [1, 2]. The partial or complete lack of blood flow to a certain part of cerebral region leads to death of the brain cells in that region. The brain cells do not die instantaneously or simultaneously but rather in a progressive manner, the rate of which, as well as the number of affected brain cells, depend on the location and severity of the interruption in blood flow [2]. For most cases, the brain cells start to die on the order of minutes and if the interruption of blood flow is longer than a few minutes, the resulting brain damage is generally irreversible [15]. The death rate and number of the brain cells lost are the primary factors that determine the outcome of a stroke event. The result of a stroke event can vary from transient numbness in a small part of the body to hemiplegia/paraplegia or death. The most

commonly observed symptoms of stroke are impairment of speech, limitation of general movement abilities such as paralysis and loss of sensation in certain parts of the body, especially face and extremities [16]. Mild to severe headache is also reported as a frequent accompanier to these symptoms.

2.1.2. Stroke types and mechanism

Stroke can be analyzed in two main categories: ischemic stroke and hemorrhagic stroke [17]. Ischemic stroke is the interruption of the blood flow partially or completely due to a blockage in a blood vessel, namely vascular stenosis or occlusion, respectively [5]. Blood clots are the most common cause of these blockages, which are called emboli although fat globules, gas bubbles or other foreign material may also be the blockage reason (embolus). A blood clot or coagulum is the semisolid material formed mostly by blood platelets as a result of clotting of blood. The clotting or coagulation process involves activation, adhesion and aggregation of platelets, with a blood protein called Fibrin also contributing to the loss of liquidity. Blood clotting is a natural phenomenon that normally occurs when a vein or artery in body is ruptured. It can be considered as a defense mechanism of the body to prevent the pooling or loss of blood depending on whether the rupture is internal or external, respectively. While the normal process of blood clotting helps the survival of the body, unnecessary or inappropriate clotting of blood might pose a great danger as the resulting clot might set free and travel through the circulatory system, eventually reaching a vascular region that is too narrow for it to pass thus blocking the blood flow in that region. If such blockage occurs in a cerebral artery, the patient is said to have had an ischemic stroke or cerebral ischemia attack.

Atherosclerosis, which can be described as the presence of fatty deposits called atherosclerotic plaques lining the vascular walls increase the risk of occlusion and stenosis as it causes a reduction in the effective cross sectional area of blood flow. For severe cases, occlusion may even occur without any emboli as the heavy accumulation of atherosclerotic plaques would be enough to block the artery completely. Arterial occlusion or stenosis are considered as the major causes of ischemic stroke while the blockage of a cerebral vein might occasionally cause an ischemic stroke as well [18]. Furthermore, a small portion of ischemic strokes are reported to originate from a general decrease in blood supply without significant occlusion, such a shock [19].

It is estimated that around 80% of the strokes are ischemic stroke events [6]. Despite these events are all based on occlusion or stenosis of a cerebral vein, the exact cause of many ischemic stroke events could not be understood. These are called cryptogenic ischemic attacks and constitute 30-40% of all ischemic stroke events [20].

A number of classifications of ischemic stroke events exist in literature. Some of these classifications, like Oxford Community Stroke Project classification (OCSP) are based on the initial symptoms of the stroke while classifications focusing on the origin of the stroke event also exist. For instance, TOAST (Trial of Org 10172 in Acute Stroke Treatment) performs classification based on the arterial size and origin of embolism [21]. One of the most widely recognized method of classifying the ischemic stroke events is classification by temporal course [22]. In this notion, the severity of the ischemic attack is determined according to the duration and reversibility of the symptoms. Hacke et. al. suggest that it is reasonable to distinguish the cases in which the symptoms are completely reversible within 24 hours than the other forms of ischemic attacks [23]. These types of ischemic attacks are known as transient ischemic attacks or TIAs. The symptoms of TIAs last only for a short time, usually on the order of minutes with less than 10% of TIA cases are reported to have symptoms lasting longer than 6 hours [24]. The most common symptoms of TIAs are motor and sensory defects, aphasia and loss of vision, generally in one eye [17].

Approximately a quarter of the ischemic stroke events are reported to be TIAs. Despite their seemingly mild nature due to the patients' full recovery within a short time, TIAs are still considered as a serious health condition as 17% of patients were reported to have an ischemic attack within the 6 months follow-up of TIA with a 12,3% mortality rate at 1 year [25]. For this reason, TIAs are also regarded as a predictor of subsequent stroke and stroke-related deaths [25].

Hemorrhagic stroke events are triggered by the rupture of a cerebral blood vessel, resulting in pooling of blood in brain tissue. This pooling results in a compression in the brain tissue, damaging the brain cells. This event is considered as an intracerebral hemorrhage, commonly known as a cerebral bleed. Most intracerebral hemorrhage events are due to hypertension because of its deteriorating effect on the blood vessels' walls. The risk of intracerebral hemorrhage is shown to be greater if the hypertension is chronic and poorly controlled [26]. Although generally considered to be more violent and deadly when

compared to ischemic stroke, it is shown that a hemorrhagic stroke survivor can recover with fewer deficits compared to similar-sized ischemic infarcts should the hemorrhage be properly resorbed [26]. The outcome of the hypertensive hemorrhage is likely to be graver if the hemorrhage extends to posterior fossa or ventricular system [27].

There are two main structural deformities that increase the risk of rupturing of a cerebral blood vessel; namely cerebral aneurysm and arteriovenous malformation (AVM) in brain. A cerebral aneurysm is the enlargement of a section in a cerebral blood vessel due to the pressure of the blood flow, forming a balloon-like shape before rupturing at some point. AVM, on the other hand, is the formation of abnormal linkage from an artery to a vein, bypassing the tissue that the artery is supposed to provide blood to. Cerebral aneurysms and AVM are mostly congenital malformations and in their presence a hemorrhage is more likely to occur in these structures than anywhere else in the body. Although such hemorrhage may also occur due to hypertension or other illnesses, a considerable portion of aneurysm and AVM related hemorrhages are reported to be traumatic and spontaneous [28]. While presence of AVM or cerebral aneurysms is known to increase the risk of hemorrhagic stroke considerably in case of a head trauma, having a head trauma is rather a random event. Furthermore, an expert system can not possibly estimate the risk of a person having a head trauma due to a fall or traffic accident from his/her medical data. Considering these facts, AVM and aneurysm related cases are excluded from this study in order to prevent a possible bias due to their random nature.

Hemorrhagic stroke can also occur following an ischemic stroke event, especially if the affected area due to ischemia is large [29]. In this case, the hemorrhage occurs in the infarct region and is associated with the earlier ischemic stroke. It should be kept in mind that the hemorrhagic stroke should have occurred within 24-hour follow-up of an ischemic stroke in order to be considered to be related to the ischemic event confidently. Vasculopathy, decrease in cerebral perfusion pressure and mass effect of ischemic edema developing after an ischemic stroke are considered as the possible reasons of the subsequent hemorrhagic stroke [30]. Lastly, according to Molina et. al., early reperfusion of ischemic tissue in brain also increases the risk of hemorrhagic stroke [31].

2.1.3. Stroke facts

Stroke is a health condition that is one of the leading causes of death worldwide. It is observed in all communities with slight variations in incidence or prevalence rates. The mortality and long term dependence statistics are somehow more dependent on the socio-economical status of the population in question as well as the quality and accessibility of healthcare services. In this section, important statistical information regarding stroke events is presented as using reliable and consistent stroke data is considered crucial in any application or study regarding stroke [10].

The number of new stroke cases worldwide was reported to be 13 676 000 by 2016 while the number of living people who had had a stroke was estimated to be more than 80 000 000. By projection, it is estimated that a quarter of people worldwide has had or will have a stroke attack during their lifetime [32]. By 2020, 795 000 people in the United States of America (USA) were reported to have a stroke, implying that a stroke attack occurred roughly every 40 seconds in the country.

92% of all stroke events in the world were reported to occur in patients above 44 years old. Moreover, the mean age of first-time stroke patients was given as 75 in western countries [33-35]. These facts indicate that there is a strong relationship between stroke risk and age.

A slight handicap towards men is apparent as it is estimated that around 52% of stroke cases occurred in men. 53% of stroke related deaths were attributed to men. When deaths per capita are considered, this means that men have a 4,1% higher risk of dying from a stroke than women. Moreover, 56% of healthy life loss, either by death or long term disability, was also accounted for men. Therefore, it can be concluded that men are under a greater risk both by means of having a stroke and having a graver outcome.

Ischemic stroke events constitute the majority of worldwide stroke attacks as 9 556 444 ischemic strokes were reported in 2016, accounting for 70% of all stroke events within that year. When compared to statistics of previous years, a slight decrease in the rational incidence of ischemic stroke is observed. On the other hand, the number of living people who had experienced an ischemic stroke is reported to be over 65 000 000, accounting for

81.3% of worldwide stroke survivors indicating that the aforementioned ratio was higher in the preceding years.

Men are more prone to ischemic stroke as 52% of new ischemic stroke cases were reported to be in men. However, when death rate is considered, despite experiencing ischemic stroke events more frequently, men might be thought to have a slightly higher chance of surviving one as 51% of ischemic stroke related deaths were accounted to women in 2016. It should still be kept in mind that this is the case for a single year and when long-term survival rates considered, it is seen that men account for 49% of the worldwide stroke survivors. Therefore, it can be concluded that men have a relatively lower chance of long time survival in the aftermath of an ischemic stroke event.

The incidence of hemorrhagic stroke events was 4 120 300 in 2016, accounting for roughly 30% of all stroke cases. On the other hand, the number of living people who had had a hemorrhagic stroke event was slightly over 15 000 000, corresponding to only 18.7% of worldwide stroke survivors. What is more, the number of deaths due to hemorrhagic stroke events were reported to be 2 838 062 in 2016, which is more than 51% of annual deaths due to all stroke events. Considering this fact, it is observed that hemorrhagic strokes are much more deadly than ischemic strokes as the risk of death due to a hemorrhagic stroke is roughly 2,4 times higher than that of an ischemic stroke attack.

When the distribution of hemorrhagic stroke events among sexes are analyzed, it is observed that 53% of hemorrhagic strokes occur in men, indicating a more visible unbalance when compared to general or ischemic stroke events. Moreover, the hemorrhagic stroke related deaths were also more frequent in men as 56% of such deaths were attributed to them in 2016. Long term survivability after a hemorrhagic stroke is reported to be lower for men as 51% of the living people who had had a hemorrhagic stroke event were women. Here, it can be concluded that men are generally under a greater risk of having a stroke event, as well as losing their lives after having one for both stroke types.

2.1.4. Stroke risk factors

Stroke is a health condition that involves different complex mechanisms taking place for different stroke types. There are some known risk factors that were proven to affect the risk

of stroke such as hypertension, dyslipidemia, diabetes or smoking. These factors are known to increase the risk of all stroke types, although some might have a higher impact on the risk of one stroke type than another. Apart from these "conventional" risk factors, there are many more risk factors that were either newly discovered or yet to be proven by extensive studies. Such risk factors are still widely accepted and considered by medical professionals in general [36]. For this reason, important parameters that are widely recognized as stroke risk factors in literature are provided under this section while it should be kept in mind that other factors were also considered in this study, which are discussed in Chapter 3: Dataset.

Hypertension

Hypertension is the health condition in which the blood pressure is generally high over extended periods of time. Diagnosis of hypertension is generally verified when both systolic blood pressure (SBP) and diastolic blood pressure (DBP) readings are over 140 and 90 mmHg, respectively. Hypertension is categorized in two, depending on the underlying cause. In primary hypertension, which makes up to 95% of all reported hypertension cases, there is no known cause of high blood pressure and it usually progresses slowly over time. In secondary hypertension, however, there is an identifiable cause of the condition such as kidney failure, congenital blood vessel abnormalities or drug usage [26]. The progress of secondary hypertension is much more rapid and the levels of SBP and DBP are higher when compared to primary hypertension. Secondary hypertension cases are much less common as they are observed in only around 5% of the population.

Hypertension is considered as the single most important risk factor in stroke with more than 57% of stroke cases in 2016 were reported to have hypertension [10]. Hypertension might trigger different mechanisms of stroke, increasing the risk of both ischemic and hemorrhagic attacks. Formation of atherosclerotic lesions and local emboli due to hypertension are considered as the most common causes of ischemic stroke while hemorrhagic stroke can occur with rupturing of a cerebral artery due to high blood pressure.

Hypertension has a well proven relation with stroke events. Framingham suggested that stroke risk increases exponentially with the degree of hypertension. Even mild cases of hypertension such as stage 1 (140-159 mmHg SBP) were shown to increase the stroke risk by 50% when compared to patients without hypertension. The risk of stroke increases up to

300% for cases diagnosed with heavier hypertension such as stage 2 (160-179 mmHg SBP) or stage 3 (SBP > 179 mmHg). In the study, the effect of patient age was also observed and it is shown that the effect of hypertension demonstrates as slight decrease for older patients. For patients aged 80-89 years, the increase in stroke risk was around 100% for stage 3 hypertension cases while the same level of hypertension was shown to cause more than 300% increase of stroke risk for subjects between 50-59 years. In the study of Franklin et. al., stroke mortality rates showed a similar trend by increasing considerably with systolic and diastolic blood pressure levels of subjects [37], as given in Figure 2.1.

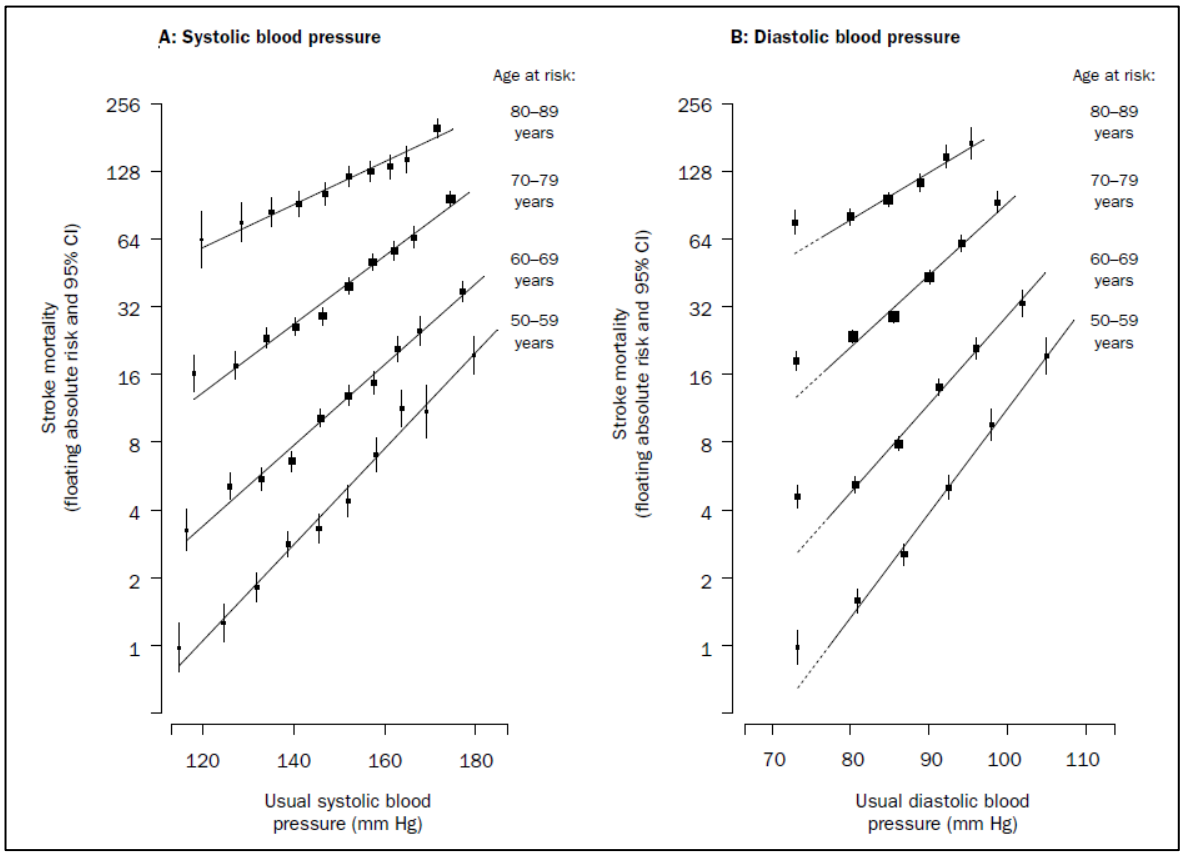


Figure 2.1. Stroke mortality rates with respect to age, systolic and diastolic blood pressure

Diabetes mellitus

Usually referred to simply as diabetes, Diabetes mellitus (DM) is a metabolism related health condition that is defined by chronically high blood sugar level which is generally caused by the deterioration in blood sugar control mechanism of the body. There are two main types of DM classified according to the origin of the disorder.

Type 1 DM results from the lack of insulin, the main hormone responsible for lowering the blood sugar level, in the body. The cause of insufficient insulin is pancreas failure due to the loss of beta cells which are the cells that synthesize insulin [38]. The lack of insulin disturbs the blood sugar balance, resulting in elevated blood sugar levels that can be treated by supplying the deficient insulin to the body externally. For this reason, type 1 DM is also known as insulin dependent diabetes mellitus (IDDM). Juvenile diabetes also corresponds to this form of DM as the symptoms of type 1 DM usually arise at an early age without a known underlying reason.

In type 2 DM, the insulin levels in body are usually within normal limits, however due to an alteration in cell operation, the body cells become less responsive to insulin, failing to adjust the blood sugar level properly. Insulin insufficiency may or may not occur in later stages of type 2 DM. Insulin resistance which is likely to develop due to obesity and lack of exercise is considered as the main cause and premise of type 2 DM. Gestational diabetes, which occurs during pregnancy might also develop to type 2 DM. However, 90% of the time it dissolves after giving birth and therefore is not considered as a separate type of DM [39].

Despite it is shown by numerous studies that both types of DM increase the risk of ischemic stroke, the mechanism is not fully understood. However, it is suggested that the ischemic stroke risk in diabetic patients increase due to deformations in cerebral blood vessels, possible platelet aggregation and impaired fibrinolysis caused by DM. It is also revealed that type 1 DM has a stronger relation with elevated ischemic stroke risks than type 2 DM. There is, however, a wide variety of results regarding how much this risk is increased by both types of DM. The claimed relative increase in ischemic stroke risks range from around 600% to 800% for type 1 DM [40] and from 50% to 500% for type 2 DM according to different studies in literature [41-45]. Due to its progressive nature, the risk of ischemic stroke in patients with type 2 DM is reported to increase with the duration of the disease [46].

Type 1 DM is also associated with an increased risk of hemorrhagic stroke, primarily due to increased levels of hemoglobin A1c [47]. The relation of type 2 DM with hemorrhagic stroke risk, on the other hand, is more disputed. Some studies conclude that there is a positive association between hemorrhagic stroke risk and type 2 DM [48] while others argue that there is no significant relation between them. There are also a number of studies suggesting that type 2 DM actually decreases the risk of hemorrhagic stroke [49].

Dyslipidemia and cholesterol

Dyslipidemia is the medical condition characterized by presence of abnormal levels of lipids in blood which are commonly referred to as Cholesterol. While dyslipidemia includes both hyperlipidemia and hypolipidemia, corresponding to excessively high and low levels of lipids, respectively, it is generally used interchangeably with hyperlipidemia as a vast majority of dyslipidemia cases are reported to be hyperlipidemic [50].

Presence of high cholesterol in blood, especially low density lipoprotein (LDL) cholesterol is known to increase the ischemic stroke risk due to plaque build up in blood vessels. If the plaque build up takes place in cerebral blood vessels, occlusion or stenosis may directly cause an ischemic stroke event. Blockage of a cerebral vessel due to a thrombus formed by such plaques formed elsewhere in body is also possible, which will similarly cause an ischemic attack. Having low levels of high density lipoprotein (HDL) alongside high LDL is shown to further increase the stroke risk while there are conflicting, though mostly towards positive, results regarding the relation between total cholesterol and ischemic stroke risk [51-53]. In more recent studies, it is suggested that total cholesterol should be considered alongside LDL and HDL levels [54]. Tirschwell et. al. showed that there is a strong link between a high total cholesterol to HDL ratio and ischemic stroke risk [55].

The relation between cholesterol and hemorrhagic stroke risk, on the other hand, was studied to a lesser extent and in most studies the relation was found out to be negative. For hyperlipidemia cases, the decrease in hemorrhagic stroke risk with respect to total cholesterol was relatively weak [56,57] and in some studies the risk of hemorrhagic stroke is reported to increase with serum cholesterol [58]. On the other hand, for hypolipidemia cases, a fair increase in hemorrhagic stroke risks were observed in the approximately 70 000 subject study of Eastern Stroke and Coronary Heart Disease Collaborative Research Group [59].

Triglycerides are an important type of blood lipids, molecules of which are formed by the merging of 3 fatty acid and 1 glycerol molecules via ester bonds [60]. The mechanism and effect of triglycerides on stroke risks are known to be similar to that of cholesterol [61].

Cardiac risk factors

Disorders associated with cardiovascular system, especially the heart are likely to result in an increased risk of thrombus formation, thus having an ischemic stroke. Although some risk factors such as hypertension, which may originate from certain cardiac disorders, are both related to hemorrhagic and ischemic stroke, the direct effect of cardiac disorders on stroke risk is almost exclusively the elevated risk of embolism. Therefore, cardiac risk factors are generally associated with ischemic stroke.

Arrhythmia is presence of irregularity in the rhythm or pace of heartbeats [60]. Beating of the heart is regulated by electrical impulses triggering the cardiac muscles. Disruptions in the properties of these signals result in a deviation from normal heart rate and pattern. There are four main subtypes of arrhythmia, depending on how the characteristics of abnormal heart beating.

- Tachycardia is the beating of heart faster than 100 bpm in rest. It is usually used to describe sinus tachycardia, in which the beating pattern is normal. Other cases where a disturbed pattern alongside tachycardia is observed, are usually referred to as atrial fibrillation and flutter.
- Bradycardia is the beating of heart slower than 60 bpm in rest.
- Atrial fibrillation is caused by random, chaotic electrical pulses that disturb the normal pattern of heartbeat. The disturbance is in the form of mild to severe tachycardia as well as a considerably irregular beating pattern.
- Atrial flutter is caused by a single abnormal electrical pulse transversing the atrium and results in an increased beating rate as well as a disturbed rhythm. However, the overall pattern is closer to normal, when compared to atrial fibrillation.

Sinus tachycardia has no direct known effect on stroke risk. However, it is suggested that patients with sinus tachycardia are more likely to develop atrial fibrillation and flutter, which are considered one of the most important ischemic stroke risk factors [62]. There is no evidence that bradycardia is related to a possible stroke event in literature, therefore it is not considered as a stroke risk factor unless reported alongside other types of arrythmia. Despite the characteristics and intensity of the causing electrical signals are somewhat different,

atrial fibrillation and flutter are usually considered together as their effect on stroke risk is common. The consecutive and irregular impulses to heart muscles cause the contractions to become irregular and rapid but also weaker. Due to weaker contractions, the amount of blood pumped to arteries is decreased, consequently causing a decrease in blood flow rate in the atrium. This results in pooling and stagnating of blood which increases the risk of clot formation greatly. A clot formed this way might cause the blocking of cerebral blood vessels, causing an ischemic stroke. Such stroke events are also referred to as embolic stroke since the emboli is formed outside the cerebral region and caused the occlusion of a cerebral blood vessel. Thrombus formation in cardiovascular system also yields in an increased risk of myocard infarctus (MI) by possibly blocking a coronary artery.

MI is an acute event that involves occlusion or stenosis of a coronary artery, resulting in reduced or stopped blood flow to cardiac tissue. Depending on the degree of stenosis, as well as the region affected, the result of the MI can vary from arrythmia to heart failure and cardiac arrest. It should be noted that, minor MI events that result in arrythmia are considered almost as dangerous as major failures as they have a high risk of relapsing and triggering another, more serious MI. Acute MI is considered as a risk factor for ischemic stroke as the absolute risk of ischemic attack within the first month following a MI is reported to be 2% [63]. This corresponds to an adjusted (relative) increase in ischemic stroke risk by 3000% which highlights the substantial effect of MI on risk of ischemic stroke [64]. An increase in hemorrhagic stroke risk is also reported to be observed in the short term follow-up of MI by up to 2000%. While the long term risk of ischemic stroke remained persistent, with MI survivors having 60% increased risk of ischemic stroke within 30 years following the MI event, the elevated risk of hemorrhagic stroke was shown to be more transient, with relative risk decreasing to 10% within the same period of time. It is argued that the increase in hemorrhagic stroke risk is not directly because of MI event, but instead due to the anticoagulative treatment following it, which is the principal procedure [65].

Infective endocarditis is the bacterial infection of the inner region of the cardiac tissue (endocardium) and/or the heart valves. One of the most common complications of infective endocarditis is embolism, which is reported to occur in almost half of the cases [66]. The risk of emboli formation is reported to be higher for cases with heart valve infection, up to 82% [67]. The structural alteration of cardiac valves due to the infection may yield an inflammatory reaction in the form of thromboembolic complications, increasing the risk of

thrombus formation considerably [68]. Therefore, it is suggested that there is an increased risk of subsequent ischemic attack due to septic emboli, especially during the first two weeks of the infection [69]. Ischemic stroke attacks are observed 60 to 120% more frequently in cases with infective endocarditis according to various studies in literature [70-72]. The mortality rates of infective endocarditis related ischemic attacks are also reported to be higher than that of ischemic attacks from other etiologies, approaching up to 40%. [73].

Valve replacement is a surgical operation that is carried out to replace a faulty heart valve with a mechanical prosthesis. Its association with stroke risk is the increased risk of embolism due to imperfect operation of the prosthesis, especially for the cases with prosthesis–patient mismatch (PPM), of the prosthetic valve. This usually takes place as a thrombotic prosthetic valve occlusion; the risk of occurrence is reported to be 1 to 20% depending on the type of the prosthesis. [74-77].

In order to reduce the risk of thrombosis, mechanical prosthesis recipients are almost always given anticoagulative treatment, which increases the risk of hemorrhagic stroke, especially in elderly patients [78]. Another secondary effect of mechanical prostheses is the elevated risk of infective endocarditis due the foreign valve surface and sewing ring, which contribute to the increase in ischemic stroke risk [79]. A similar conclusion could also be drawn for pacemakers, the leads of which were observed to cause emboli formation in 30% of the cases, some portion of which resulted in ischemic stroke [80].

Smoking

Smoking is considered as a behavioral risk factor that increases the stroke risk by causing mostly harmful changes in the body in different ways due to the toxicity of chemicals. It is shown to be associated with increased levels of lipids, especially triglycerides in blood, which may cause plaque formation in blood vessels thus increasing the ischemic stroke risk [81]. Smoking is also associated with decreased HDL levels, the effect of which was discussed under Dyslipidemia and cholesterol.

Some chemicals in tobacco smoke are known to cause physical alterations on the surface of blood platelets, which makes them more prone to unnecessary clotting [82]. This increases the risk of a thrombus formation, thus increasing the risk of an ischemic stroke event. Toxic

effect of chemicals contained in tobacco smoke includes loss of elasticity in blood vessels, increasing the risk of rupture and therefore hemorrhagic stroke as well. The relative increase in risk of hemorrhagic stroke is reported to be higher than that of ischemic stroke, by a factor of 1,5 [83].

Migraine

Migraine is considered as a neurovascular disease that involves episodes of headache and autonomic nervous system dysfunction in patients. Predominantly characterized by pulsating unilateral headaches, migraine episodes usually last for 12 to 72 hours. In some cases, nausea might also accompany the headache attacks. These are called vestibular migraine cases. In around 67% of migraine patients, the attacks have little to no accompanying disorders other than sensory sensitivity and blurred vision. These cases are called migraine without aura (MwoA). Migraine with aura (MwA), on the other hand, is characterized by additional symptoms called auras, usually start before the headache episodes. Auras may also persist during the attacks in some cases. Most auras are of visual type, ranging from blind spots to flashes of light or tunnel vision. Confusion, muscular spasms, auditory hallucinations and speech problems are among rarer types of auras. In some patients, auras may not be present in every single migraine attack. These patients are said to have both types of migraines. Rarely, the migraine attacks may in the form of auras only, without a headache or nausea. These are called silent migraine cases.

In recent studies, previously unknown association of migraine attacks with long term dysfunction of cerebral arteries were discovered [84]. Although the mechanism is not fully understood yet, observation of infarct-like brain lesions in migraine patients has led to a hypothesis that migraine patients have a higher risk of developing ischemic stroke [23]. This hypothesis is strengthened by the outcome of the meta-analysis carried out by Etminan et. al. which suggests that patients with MwA and MwoA have 127% and 89% higher risk of having an ischemic stroke when compared to patients without migraine [85].

Drugs

There are a number of drug types, generally based on their active ingredient that are known to have an effect on stroke risk. Although having a broad range of main objective, drugs that

affect the components of circulatory system, sometimes unintentionally, are very likely to affect the stroke risk as well.

Statins, a group of lipid-lowering drugs, were shown to have a decreasing effect on ischemic stroke risk by up to 20% [86]. However, as examined by SPARCL trial, the usage of statins is also associated with an increased risk of hemorrhagic stroke [87]. This was considered as an expected result that could be explained by the negative relation between cholesterol and hemorrhagic stroke.

Oral contraceptives (OC), commonly known as birth control pills, are the drugs used by women to avoid pregnancy. Mainly consisting of estrogen and progestin, they act as externally taken hormones to prevent a possible pregnancy by altering the levels of these hormones in body, eliminating the suitable conditions for pregnancy. The effect of these hormones, when taken in the form of OC, are not limited to preventing pregnancy though. It is proven that OCs also affect coagulation and fibrinolysis, increasing the risk of thrombus formation. There is some evidence that OCs might cause temporary hypertension or alter the lipid and carbohydrate mechanism. These effects are shown to result in an increased risk of stroke by a factor of 2 to 5 with the usage of OCs [88].

Anticoagulants or blood thinners are drugs used to decrease the coagulating property of blood by inhibiting the blood clotting factors such as fibrinogen, prothrombin or thromboplastin. Anticoagulants are generally used as a precaution against an increased risk of clot formation within body which may be due to a recent cardiovascular surgery, an ischemic attack or cardiac disorders. Their effectiveness in reducing the risk of ischemic stroke by reducing the risk of thrombosis was well proven with numerous studies [89,90]. On the other hand, anticoagulants' interference with clotting process was shown to cause an increase in the risk of hemorrhagic stroke events by up to 22% [91]. Acetylsalicylic acid (Aspirin) is reported to have a similar effect due to its anticoagulative properties and therefore, is also associated with an elevated risk of hemorrhagic stroke [92].

The term drugs may also refer to illegal substances with intoxicating effects that are used for recreational purposes. Drug abuse is considered as the usage of such intoxicating drugs, as well as the misusage of conventional drugs, unrelated to their intended medical purpose. It is suggested that illicit drug usage is another factor increasing the risk of stroke, especially among young adults [93].

Cocaine, which has an approximately 4500 years of history is considered as an illegal drug in most parts of the world since early 20th century. It affects the central nervous system towards euphoria and during this process, blood pressure and heart rate is elevated. For this reason, cocaine usage is associated with increased risk of stroke, especially hemorrhagic stroke as 80% of stroke events observed in cocaine addicts were reported to be of hemorrhagic type [94].

Amphetamines, which were initially used as bronchodilators in medical literature are among illegal drugs today and their effect on increased stroke risks are studied since 1945 [95]. Although they correspond to a wider range of compounds, amphetamines are mostly used in the form of meth (Methamphetamine) and ecstasy (MDMA). In recent studies, usage of amphetamines was shown to increase the risk of hemorrhagic stroke attacks while no observable effect on ischemic stroke risks were reported [96].

Heroin is an addictive drug, synthesized from morphine and is illegal in most parts of the world, excluding England, where it is considered as a prescription analgesic drug. The relation between heroin and increased stroke risks were discussed in many studies, suggesting a considerable increase in risk of both ischemic and hemorrhagic stroke attacks due to heroin abuse, especially within first 12 hours of intake [97-100].

Cannabis, which is used as a source of fiber in rope industry is a plant, farming of which is restricted due to the psychoactive chemicals it produces. These are known as cannabinoids and can be taken into body in the form of Marijuana, which is an illegal drug with addictive properties. Considered as a “lighter” drug, its overall effect on body is relatively milder when compared to drugs such as heroin or meth. Its relationship with increased stroke risk is not as clear as other drugs discussed, however, it is suggested that cannabis usage increases the risk of stroke through arterial spasm, atrial fibrillation and flutter which are common effects of Marijuana [101,102]. There are other illegal drugs such as LSD the relation with stroke risks were not completely clarified with some conflicting studies [103-105]. Due to patients’ reservation in admitting using such illegal drugs, it is often difficult to obtain reliable data of drug usage and therefore they are only occasionally considered in cohort studies and clinical datasets.

Alcohol

The effect on alcohol consumption on stroke risk is relatively complex, as it is impossible to define a general rule without considering the amount of alcohol intake, as well as the type of stroke.

Alcohol is known to cause alterations in plasma lipoproteins with the most remarkable effect being increased levels of HDL cholesterol. Moderately high HDL cholesterol is considered desirable as HDL particles cohere to cholesterol in blood to carry them back to liver, lowering the total cholesterol in body tissues, as well as its accumulation on arterial walls. From this perspective, alcohol consumption could be claimed to have a positive effect on reducing the risk of ischemic stroke. A large number of studies suggest that light (up to 20 g/day) to moderate (20-60 g/day) alcohol consumption actually decreases the ischemic stroke risk, as well as mortality rates, by up to 20% [106-109]. Wine is also shown to further decrease the risk of ischemic stroke thanks to its polyphenols that have anti-thrombotic effects [106].

Although having a decreasing effect on ischemic stroke risk with light to moderate consumption, alcohol is shown to increase the risk of ischemic stroke with heavy consumption (>60 g/day) [110]. This phenomenon is mainly due to alcohol's effect on cardiovascular system, such as causing atrial fibrillation and hypertension while its contribution to elevated lipid levels in blood should also be considered.

The relation of alcohol consumption and risk of hemorrhagic stroke is considered to be relatively straightforward. Alcohol intake has two important effects in body in terms of hemorrhagic stroke risk. Firstly, it increases the blood pressure, which is shown to be one of the most important risk factors of hemorrhagic stroke. Secondly, alcohol consumption is shown to disrupt the coagulation process, causing diminished fibrinolysis as well as impairing the platelet function. Heavy consumption of alcohol also damages the liver, reducing the platelet count in blood as well as many other adverse effects on circulatory system. An impaired coagulation process is reported to be one of the main causes of hemorrhagic stroke alongside increased morbidity [20]. For these reasons, it is suggested that hemorrhagic stroke risk increases with alcohol consumption in a roughly linear fashion.

In this section, a number of important risk factors were presented alongside their mechanism on ischemic and hemorrhagic stroke events. Even though a longer list could be presented, a complete list of stroke risk factors is yet to be formed. World Stroke Organization (WSO) discussed that all known risk factors of stroke combined account only for 88% of the reported stroke cases, implying that 12% of stroke events occur in the "perfectly healthy" portion of the population [10]. Although a small fraction of such cases could be relied on undetected or unreported factors, it is self-evident that there must be other stroke risk factors that are yet to be discovered. For this reason, a large number of possible risk factors are considered alongside well known ones when preparing the dataset to be used in this study.

2.2. Engineering Background

In this study an expert system employing various algorithms of machine learning is proposed. In order to provide a better understanding of the development process as well as the operational principles of proposed model, basic principles of artificial intelligence and artificial neural networks must be understood first. Moreover, comprehending principles of different types of machine learning are necessary to use them effectively while designing a new model. A number of important machine learning methods are also briefly presented to familiarize the reader with the models formed in the study. Finally, the properties of expert systems are explained alongside their functionality.

2.2.1. Artificial intelligence

Artificial intelligence (AI) is the form of intelligence that employs intuitive approaches to perform mental tasks such as inferring, decision making or recognition instead of definite numerical algorithms. By employing such approaches, it is aimed to bring the inference process closer to mechanism of human mind [111]. Artificial intelligence methods involve the imitation of human intelligence through observations towards a specified goal [112]. The imitation is not only limited to human thinking but also includes approaches adopted by humans in problem solving.

Although artificial intelligence is considered almost exclusively in association with digital computers or computer controlled robots, its principal ideas were set much earlier, in prehistoric ages [113]. This very early notion of artificial intelligence was presumably

different than what is used today. However, the principles of symbolic AI, that is the branch of AI that relies on the usage of a physical symbol system in order to perform intelligent action [114] were already set and used, though not explicitly defined, in Ancient Greece [115]. Symbolic AI systems are generally based on formal logic, which involves representation of propositions, statements and deductive arguments by symbolic structures [116]. In symbolic AI, the inference process is carried out by the mechanical manipulation of these symbolic structures which was possible before the digital computers. The premise of symbolic AI is considered to be Pythagorean number theory, which studied the numbers through arrangement of pebbles. It is no coincidence that the Latin word for pebble is “calculus”. Pythagoreans were able to classify the numbers according to the shapes they could obtain using the pebbles. For example, the term square number were used to describe the numbers which could be represented by arranging the pebbles in such a way that they are equally spaced, both horizontally and vertically. This shape was indeed a square, which is still used in mathematics to express the second power of a number. The usage of symbols in number theory to suggest, explain and prove theorems paved the way to the usage of calculi (pebbles) not only for computation, but also for reasoning.

Pythagoras suggested that relative consonance of pitches in music could be rendered logical and rational by reducing it to numerical ratios [117]. This was considered as an important example of demonstrating an otherwise intuitive phenomenon could be reduced to a formal structure that could be displayed and calculated using symbols. In parallel with the discovery of such methodology, Pythagoreans realized that almost everything could be reduced to numbers or symbols and therefore made intelligible, rational and logical [118]. Consequently, this meant all knowledge could be expressed in terms of a systematic arrangement of otherwise meaningless symbols. Since the knowledge, or information could be learned and used, it was straightforward to assert that the arrangements representing knowledge had the same properties as well. This highlighted the fact that artificial intelligence was a possible, realizable notion. It would however take a couple of millennia until the first digital applications of artificial intelligence could be implemented.

Modern AI, which is considered to emerge with Turing’s 1950 paper "Computing Machinery and Intelligence" is considered as a branch of computer science [115]. AI based systems employ digital computation methods to produce results or perform a task, which is also a property of any other computerized system. The distinctive property of AI is that it performs

such tasks without the necessity of a user indicating how to. This means that, after the training process, an AI based system is able to perform the necessary task by making inferences, judgements and decisions on its own. The mechanism of these processes was created during the training stage by the designer by providing examples to the AI. This training procedure may or may not involve the guidance of the designer also providing the expected outputs. In either case, the intelligent capability of AI relies on the examples it was provided instead of a set of definite instructions, which is the case for conventional computerized systems.

2.2.2. Artificial neural networks

Artificial neural networks (ANN) are an approach in AI research that aims to reproduce human-like intelligence by using artificial neurons as their core elements [112]. These artificial neurons are actually mathematical models that were created by mimicking the structure and operation of biological neurons. For this reason, in order to understand the operational principles of ANNs, it is necessary to know the basic properties of a biological neuron.

Neurons are considered as the building block of the biological neural networks in humans and animals. The human brain contains 86 billion neurons on average [119]. A typical neuron has three important components relevant to their operation: soma, which is considered as the cell body, axons and dendrites. An illustration of a biological neuron is given in Figure 2.2.

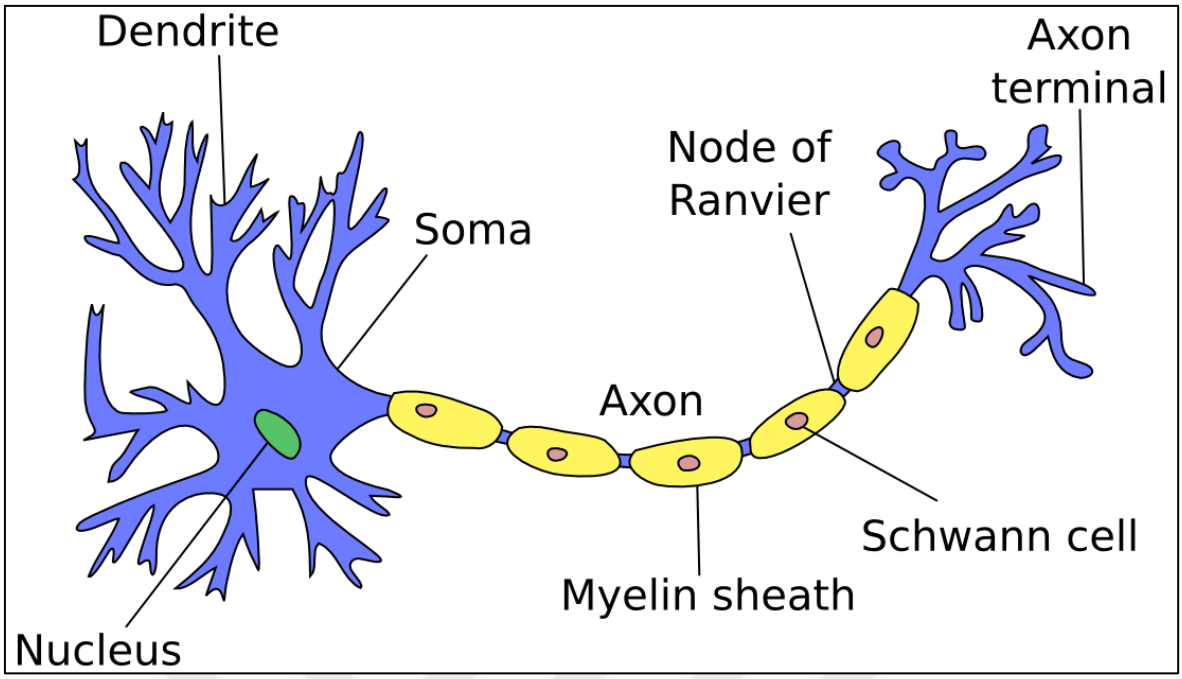


Figure 2.2. A sample illustration of a neuron

Neurons operate in a cascaded manner with other neurons, as well as the effector cells which are the target non-neural cells in body that are the final destination of the signals. Electrical and chemical signals, which are the form of transmission in neural networks are received by neuros at Dendrite terminals and transmitted via Soma to Axon terminals. The gap between the axon terminal of the transmitting neuron and dendrite terminals of the receiving neuron (or an effector cell if the transmitting neuron is terminal) is called a synapse. Synapses can either be electrical or chemical, depending on the mechanism that the signal relaying is carried out [120]. It is important to note that although electrical or chemical signals can be transferred via synapses, the signals flowing through neurons are always electrical. For this reason, rapid electrochemical processes that involve “encoding” and “decoding” of neurotransmitters take place in axons and dendrites, respectively if the synapses between two neurons are chemical [121].

The neurons are normally at rest and the electrical potential at the membrane of such neuron is called resting potential which is around -70 mV on average [121]. When a neuron receives a signal from its transmitting neighbor, the signal changes the membrane potential while travelling through the neuron which triggers opening or closing of voltage-controlled ion channels [120]. If the ion channels open due to the signal, the membrane potential decreases in magnitude. It should be noted that not all depolarizations result in transmission of the

signal to the next neuron. The received signal must create a depolarization that is strong enough to enable its transmission through the neuron and reach the axon terminal. This threshold for depolarization magnitude is called threshold potential or threshold action potential and is around -55 mV. This means that the depolarization must be greater than 15 mV in magnitude to increase the membrane potential to overcome the action potential and thus enable the transmission of the signal.

The neurons in artificial neural networks which are also referred to as processing units, consist of three important parts, or properties similar to biological neurons. The inputs of processing units are analogous to dendrites in a biological neuron. A processing unit may have, and usually has, multiple inputs linked to the unit via weights and biases. The "soma" or body of the processing unit is where the actual processing of input takes place. The output, which is similar to the axon of a biological neuron, is provided either as an input to the next processing unit(s) or is provided as the ultimate output of the network if that neuron is the last one in the ANN. Just like biological neurons, each processing unit can only have a single output.

The inputs to a processing unit are either an expression of data from outside of the ANN such as an image's RGB pixel values or outputs from other neurons in the network. These input variables are processed using the weights and biases to form the variable called net or activation according to Eq. 2.1.

$$\text{Net} = \mathbf{W} \cdot \mathbf{x} + \mathbf{B} \tag{2.1}$$

where **W** is the weight matrix, which is a vector if a single processing element is present, **x** is the input vector and **B** is the bias vector which becomes a scalar for a single processing element. Eq. 2.1. can also be written in summation form for a single processing unit as given in Eq. 2.2.

$$\text{Net} = \sum_i w_i . x_i + b \tag{2.2}$$

where w_i is the weight of i^{th} input x_i and b is the bias.

The activation is then passed through a function which is called the activation function to calculate the output of the processing unit with respect to the inputs provided [112]. The activation function can be conceptually associated with the threshold action potential of a biological neuron, but with an operational difference. The biological neuron's output depends on whether the threshold potential is exceeded or not and is practically invariant of the action potential's exact magnitude. On the other hand, the activation function is usually a nonlinear and continuous function that determines the output according to the exact value of the activation in modern ANN applications [111].

While the weights and biases of the processing elements of an ANN are randomly set in the beginning, in order to obtain meaningful outputs, the ANN must be trained first [112]. The training process, which is generally in the form of supervised learning for traditional approaches of ANN such as the perceptron, inputs with known (expected) outputs are provided to the network to obtain the calculated outputs [122]. The weight and bias parameters of each processing unit are then updated according to the deviation of calculated output from the expected output which is called the training error using a backpropagation method such as stochastic gradient descent. The training process continues until the training error is minimized to a user-defined value or maximum number of iterations is reached.

ANNs have gained a notable place among AI based computing systems with the advance in computer technology and subsequent exponential increase in computational power since the dawn of the 21st century [122]. They are many types of ANN models from simple perceptron to advanced deep models such as convolutional neural networks (CNN) or generative adversarial networks (GAN). Some significant ANN types are presented in Table 2.1 alongside some of their application areas. It is, however, beyond the scope of this thesis to discuss all ANN models in literature.

Table 2.1. Some important ANN models and their application areas

Model	Application Area
Multilayer Perceptron	Binary Classification, Natural Language Processing
CNN	Image and Video Processing, Robot Navigation
GAN	Image and Video Processing, AI-created Artwork
RNN (Recurrent Neural Network)	Machine Translation, Time Series Prediction

2.2.3. Machine learning

Machine learning is the application of AI algorithms with the aim of developing a solution to a problem according to the available data associated with the problem [123]. The common principle of all machine learning algorithms is to learn from experience, that is to learn by analyzing the examples related to the problem [124]. Machine learning is generally considered as a universal approach in AI-based classification and clustering problems.

Conventional approach in programming is the designer "telling" the computer how to solve a problem by giving directions at every step during the computation process [124]. Although seemingly trivial, this approach does not necessarily result in simple programs such as number sorters. It should be kept in mind that many state of the art applications of computer science such as automated process control systems, web servers or flight simulators could be designed with such approach.

Although some tasks such as identifying whether the animal in the given picture is a cat or bird is trivial for human brain, it is practically impossible to design a computer program that could perform the same task by using only the aforementioned approach [122]. There are certain tasks such as feature recognition, data mining or clustering that require the guidance of the designer instead of directions on how to execute every single step to solve the given problem [122]. Therefore, it is necessary to rely on machine learning to develop its own set of instructions employing AI methods to obtain solutions to such problems.

Machine learning concept was first presented by Arthur Samuel, in 1959 [125]. One of its very early applications was pattern recognition which was widely investigated in Nilsson's 1965 book "Learning Machines" [126]. Pattern recognition, which still is one of the most fundamental areas of research in machine learning, was the dominant task in this field well until 1980's when the machines' potential in not only in recognizing but also in predicting began to attract growing interest [127]. Today, recognition and prediction remain among popular areas in machine learning study alongside classification.

The term learning in machine learning corresponds to the gained ability of generalization which can be considered as the ability of the machine to perform "as expected" on new examples or tasks after being introduced to similar examples [123]. The examples that the

machine is shown are referred to as training data. The training data is considered as a sample representation of the universe or space of occurrences within the context of the task. Therefore, regardless of the task of or algorithms used by the machine, the training data should be able to represent its originating universe as accurately as possible by means of properties of its samples such as probability of occurrence, mean or variance [123].

Machine learning algorithms are primarily categorized according to their learning processes. There are three main types of machine learning algorithms in this context [124].

Supervised learning

Arguably the oldest and most widespread learning method, supervised machine learning algorithms involve the presentation of training data in the form of input-correct output pairs [128]. Therefore, each example in the dataset must have a corresponding correct or desired output in order to form a training dataset.

The supervised learning process consists of an iterative optimization procedure subject to some objective function [124]. The parameters of the network are initialized randomly, or according to the designer's expertise and the training data is fed to the network. The network produces outputs according to its initial parameters, which are initially mostly incorrect unless the initial parameters are very close to that of a carefully optimized system which is very unlikely. These outputs are then compared to the desired outputs via the objective function. The parameters of the system are then updated according to the rules and constraints defined for the model, trying to minimize (or maximize, in some approaches) the objective function. In order to claim that a learning process is effective, the objective function, which can also be regarded as the cost of the overall error of the system, must exhibit a decreasing behavior with each iteration [124].

Supervised learning is widely employed in classification, regression and pattern recognition tasks. Classification can be defined as the task of assigning a class to each input. In classification tasks the classes are discrete, therefore a method of discretization is generally applied before the final output of the system. Training process of an e-mail spam detector that labels each e-mail as "spam" or "safe" can be a good example for supervised machine learning on a binary classification task. In this case, the designer provides the system a

number of example e-mails and also tells which e-mail is spam and which is not. Throughout the learning process, the objective function, which is a function of correct and calculated outputs, will be tried to be minimized to minimize the classification error [124]. Even for relatively simple cases like this one, the objective function may vary considerably depending on the designers' approach. For example, an algorithm that prioritizes filtering spam e-mails over safe ones would have a different objective function than one that prioritizes ensuring that no safe e-mail is marked as spam. Different objective functions will expectedly result in different final models even with identical network structures and parameters.

Regression is similar to classification as the main aim of the task is to identify the given data [123]. However, in regression tasks the sample space is continuous as opposed to the discrete sample space of classification tasks. Therefore, there are theoretically infinite number of classes for each regression task [124]. In practice however, this means that the main aim of regression is to estimate the value of a variable (output) according to the inputs (features) provided to the system [129]. For instance, estimating the lap time of a race car can be modelled as a regression problem and a system can be trained using supervised learning to solve it. In this case the training dataset could contain input parameters such as the length of the track, properties of the car such as horsepower or top speed and the weather conditions while the output parameter would be the lap time. With sufficient data and accurate modelling, it would then be possible to estimate the lap time of a new car on a new track reliably.

Pattern recognition can be defined as the task of detecting and identifying the regularities or patterns present in the data automatically [123]. While the term pattern corresponds to visual patterns in most applications, any trend that can be expressed numerically or symbolically can be considered as a pattern. For example, a system used to estimate the seasonal rainfall in an area can employ pattern recognition methods to learn from a century's meteorological data. Supervised pattern recognition algorithms require the patterns to be explicitly defined and marked in the training data, which can be relatively tedious for humans when compared to classification or regression tasks [127]. However, the resulting output may also be proportionally more valuable depending on the application. To make a comparison, two systems that work on bone x-ray images to detect fractures can be considered. The classifier system would require only the labeling of whether the bone in each image is fractured or not. The pattern recognition system on the other hand, would require a medical expert's work

on each image to trace the fractured area to be trained with supervised learning. Although the required human effort is considerably high in the pattern recognition case, the resulting system would be able to detect the fractured region while the trained classifier could only state whether the bone in a given image is fractured or not. It should however be kept in mind that each approach has its own advantages and disadvantages and it is up to the designer to select which application would be most useful to solve a problem.

Unsupervised learning

Unsupervised learning involves presentation of the training data in terms of only inputs. The expected or correct outputs are not demonstrated to the system. In most cases that require unsupervised learning, these outputs are either unavailable or cannot be shaped into desired format easily. The machine is then expected to find a structure in the data that is not originally observable. This structure may correspond to a pattern, a hidden feature group or a set of implicit rules to associate the data points with each other depending on the application.

In supervised learning, it is relatively easy to determine whether the learning process is efficient or not since the objective function roughly aims to minimize the error in one way or another. However, in unsupervised learning since the expected outputs are not available or unclear, expressing an objective function in terms of “expected minus actual output” is impossible [124]. For this reason, the objective function must be designed as a cost function, the interpretation of which can vary greatly from one application to another. It is up to the designer to determine what kind of task the system should be able to and what the performance metrics are, i.e., what is expected from the system to be considered as efficient and accurate. The objective, or cost function can then be determined in terms of these metrics to achieve the desired system operation [127].

Clustering, which can be described as grouping the examples according to their similarity to each other, is one of the most widely studied unsupervised learning approaches in literature [130]. Clustering approaches are considered as a branch of data mining, particularly exploratory data mining as the similarity of samples are not explicitly defined and revealed by the clustering process. The assumptions on the data shape the algorithm’s approach on its seek for similarities and subsequent groups. In most clustering approaches, the objective

function of the learning process is defined in terms of similarity or proximity of samples in the same group and distance of samples in different groups as well as the average distance of samples in each group to the group centers. There are also other approaches that focus on estimated density and connectivity of the graph formed during training [131]. Clustering has seen many application areas including but not limited to search engines (grouping and sorting results), medical imaging, image segmentation and anomaly detection. Theoretically, any task that involves grouping of some data without strict labeling can be carried out with a properly designed clustering algorithm [132].

Reinforcement learning

Reinforcement learning falls between supervised and unsupervised learning by means of designer's guidance to the machine. It has, however, some distinctive properties that differentiate the learning process from the other two. Reinforcement learning aims to train the machine on how to interact with its environment subject to an objective. The possible interactions are given an award or penalty parameter by the designer and during the learning process, the action path with highest cumulative award is tried to be determined [129].

Reinforcement learning is similar to supervised learning in terms of designer providing some parameters to direct the learning process, although these parameters are not the expected outputs of the system. In practice, problems that require reinforcement learning approaches generally do not have known, exact solutions which is also the case for the unsupervised learning problems. From this perspective, reinforcement learning can be considered as an approach that tries to benefit from designer's knowledge on a problem without a universal solution that performs well regardless of the machine's environment. The machine's environment is typically modelled as a Markov decision process, which is a mathematical model that assumes the possible outcome of a decision is determined with a combination of random processes and the decision itself. Therefore, unlike supervised learning the solution that a reinforcement learning based model proposes may vary depending on its environmental parameters. Even when the machine's environment is held static, in most models, the output can be different at each trial due to the random nature of Markov decision processes.

Due to the aforementioned nature of reinforcement learning, it can be employed for models to perform various tasks requiring some way of interaction with the machine's environment. AI "players" that learn from the games they played against humans (or other AI) to develop strategies to win an opponent in games like chess or Go employ reinforcement learning to improve their "skill" in these games. Reinforcement learning is also adopted in autonomous vehicles to learn to adapt their driving according to the traffic and other environmental factors. Genetic algorithms, which are inspired by heredity and natural selection mechanisms in nature, are another popular area in AI that employ reinforcement learning techniques. Other areas that use reinforcement learning based applications include swarm intelligence and statistics.

2.2.4. Expert systems

Expert systems can be defined as computer-based systems that rely on AI methods to perform complicated tasks that require knowledge and expertise alongside inference and decision making capabilities [131]. Expert systems try to imitate the human experts by means of their approach in problem solving in a similar way that ANNs emulate the operation of biological neural networks. It should be noted that, computer programs with a long chain of "if-else" commands may also be referred to as expert systems in some works in literature. Such systems were available as early as in 1950's and were used occasionally in diagnostic applications in medicine [127]. These programs, however, lacked the usage of AI methods and therefore did not have a reasoning based decision making capabilities. Therefore, the term expert systems will be used for AI based expert systems throughout the study.

An expert system must be trained with machine learning tools in a similar way that a human expert must gain his/her expertise through various methods of learning and practice. The methodology followed in training stage depends on the aim of the expert system and the designer's preferences. Moreover, many modern expert systems consist of multiple AI models alongside definite processes such as data transformation. For such systems, two different training procedures can be followed depending on the structure and algorithms used in the system [131]. If the variables and operation of sub-models are relatively independent of each other, each sub-model can be trained separately and then brought together as a trained expert system. On the other hand, if the interaction between sub-models is

considerable, it is advisable to train the sub-models (and hence the system) as a whole which generally possesses a relatively difficult optimization problem subject to many hyperparameters. The systems exhibiting this property can be trained effectively using multitask learning methods which aim to optimize the learning processes of the sub-models simultaneously by exploiting the similarities between them. Multitask learning was shown to be considerably successful if the sub-models have many common parameters and the training data of each set is relatively small so that training sub-models individually would yield overfitting [133]. Expert systems are generally capable of training indefinite number of times without losing earned expertise from earlier training processes thanks to their knowledge-based structure. This is considered as a significant advantage over many other AI applications that "reset" their knowledge when trained with new data, losing possibly useful information obtained from earlier data.

Expert systems are generally employed in decision making tasks that involve analysis of a large dataset, which would practically be impossible for a human expert to analyze throughoutly in a reasonable amount of time. First examples of expert systems appeared in 1970's as such potential of computers began to be realized more clearly. They are considered among first practical applications of AI that proved substantially useful [122]. With their functionality they became the dominant product of AI research by early 1980's [127]. 1982 saw the first application of expert systems in designing a large-scale product, a logic synthesis program to generate logic gates to be more precise. The expert system, named as SID (Synthesis of Integral Design) was able to generate more than 700 000 logic gates which constituted 93% of the VAX 9000 Series computers' CPU [134]. The significance of SID's success was not limited to the fact that it was the first application of an expert system to develop such a complex product. Until then, the general opinion on expert systems were that despite their intelligent capabilities, they were an inferior imitation of human expertise and knowledge, simply exploiting the processing power of computers. SID, on the other hand was able to expand the rules encoded by its designers, generating software logic synthesis routines that are far more complex than these rules. In the end, the overall design exceeded the capabilities of experts themselves. This was verified not to be a coincidence by SID outperforming human designers in many more cases during development stage of VAX 9000 [135]. By this way, SID was the first example to prove that expert systems can achieve higher expertise levels than their designers with their AI-based learning and self-improvement capabilities. Today, expert systems still hold their position strongly, with many new

application areas being introduced with the developing technology. This is due to their unique capabilities discussed above, as well as their generally user-friendly interfaces designed for nonprogrammers' use.

Expert systems can be used to analyze and interpret descriptions from language based structures. Speech and voice recognition are the foremost tasks that such systems carry out. One of the earliest examples of speech recognition was Hearsay-I in 1973, succeeded by Hearsay II in 1977 [136]. These systems then evolved into a more general framework of knowledge-based expert systems with Hearsay-III in 1980 which could be used to design expert systems in various fields [137]. Recent studies include EDA, an electronic system design tool that is incorporated with a Polish speech processing system [138]. The expert system in their study employed AI techniques of natural language processing that incorporated maximum entropy models for natural concept generation.

Their ability to make predictions based on inference and understanding of the present state space made expert systems a viable option in prediction based tasks. Woolery and Grzymala implemented an expert system to predict preterm birth risk in their 1994 study [139]. More recent studies regarding expert systems' usage in predicting medical risks can be found under the section Literature Review.

In 1999, Lee and Jo proposed an expert system to observe the stock market and the predict best stock market timing. The system analyzes the conditions and dynamics of the stock market with a pattern-recognizing algorithm and then makes predictions using its comprehension of these patterns and developed knowledge base [140].

Expert systems were also employed in autonomous vehicle development projects as early as in 1990. That year, Naval Postgraduate School of Canada started a project to develop a small autonomous underwater vehicle (AUV). The expert system realized by Kwak et. al. for the AUV's mission control was able to design an optimal strategy depending on the constraints and objectives of a given mission [141]. That is, both planning and real-time mission execution tasks were carried out by the expert system. The coarse geographical data was provided to the expert system's knowledge base offline while the system used its mission control to generate online environmental database to avoid obstacles and update its local path continuously during mission execution. Sunberg et. al. proposed a multi-phase expert

control system for helicopter autorotation in their 2013 [142] and 2015 studies [143]. The expert system used in their study also employed fuzzy logic to implement phase changes. With the capability of make time-to-impact predictions, the expert system provided anti-flare maneuverability to the helicopter. The simulation results demonstrated the expert system's superiority over human pilots and indicated that the system was applicable to a broad variety of flight platforms. Onboard tests with the system installed on a TREX helicopter were also reported to be successful.

2.3. Literature Review

Estimating the stroke risk in patients is a problem that has drawn attention of researchers from a variety of professional backgrounds, dominated by medical and engineering branches. This being the case, a literature review involving both medical and engineering based studies are carried out, with most significant studies discussed in this section.

One approach commonly observed in literature is to develop a rule or scoring based stroke risk estimator system that uses statistical information to predict the stroke risk depending on the patient data. There are many variations of such systems, some of which are in the form of web applications on healthcare facilities' webpages. UCLA stroke risk estimator aims to assess the stroke risk of a patient according to a linear combination of stroke risk factors [144]. The risk factors that UCLA considers when calculating the risk of a stroke event includes gender, age group, SBP, hypertension, diabetes, smoking, certain cardiovascular disorders such as MI, coroner insufficiency or atrial fibrillation. Another similar stroke risk estimator is designed by Cleveland Health Clinic and available on their website [145]. Cleveland Clinic Stroke Risk calculator also considers risk factors such as ethnicity or previous stroke history when assessing the patient's stroke risk alongside the risk factors included by UCLA risk calculator with the exception of SBP. These risk estimators, despite covering the most common stroke risk factors, were only in the form of a self-assessment tool and no related studies formally discussing their formulation and performance was available in literature.

McGeechan et. al. proposed a statistical stroke risk estimation algorithm mainly focused on investigating the relation between retinal vessel caliber and stroke risk while considering well recognized risk factors as well [146]. In their 2009 study, the authors analyzed a 20 798

patient dataset from 6 different cohort studies to clarify the association between stroke risk and retinal vessel caliber which is not considered as a traditional stroke risk factor. The study results indicated that the width of retinal venular caliber could be associated with a higher stroke risk as 4.5% of the patients with wide retinal vanular caliber were reported to have a stroke during the follow-up of 5-12 years. This corresponded to a relative increase in stroke risk of 15% for patients with wide retinal vanular caliber. The authors further showed that the caliber of retinal arterioles was not associated with stroke as no relative increase in stroke risk is observed for patient groups with different caliber of retinal arterioles. Although no clear explanation regarding the mechanism of the increased stroke risk due to wide retinal venular caliber could be made, the authors discussed that taking this parameter into account resulted in reclassification of around 10% of the patients by means of stroke risk. With this study it was suggested that more accurate estimations of stroke risk could be achieved by considering not only the traditional risk factors but also relatively uncommon ones such as retinal vanular caliber width.

Goff Jr et. al. published a detailed report on 2013 regarding their study of preparing a new set of guidelines in estimating the risk of cardiovascular diseases [147]. Though not explicitly focused on stroke, a large variety of statistical cohort studies' results are analyzed to propose a detailed chart in estimating the risk of any cardiovascular disease (CVD) including MI, coronary heart disease and stroke taking place. The estimation system considered 15 different risk factors which are then used to form three parameters to be used in the final risk estimation formula. These parameters were:

- Baseline Survival: A statistical value depending on the race and gender of the patient.
- Individual Coefficient x Value: Weighted sum of age, lipids and blood pressure values of the patient (13 different coefficients in total)
- Mean Coefficient x Value: Statistical mean of Coefficient x Value for the given gender-race group

These three parameters were used to estimate the 10-year risk of CVD according to Eq. 2.3.

$$\text{CVDR} = 1 - S_{10}e^{(IndxB - MeanxB)} \tag{2.3}$$

where CVDR is the estimated cardiovascular disease risk, S_{10} is Baseline Survival, IndxB and MeanxB are Individual and Mean Coefficient x Values, respectively. Despite a number of example cases' CVDR were calculated in the study, a performance assessment of the system was not carried out, therefore the accuracy of the suggested approach could not be verified.

Although the aforementioned approaches to estimate stroke risks employed statistical and medical information that are widely recognized, the developed models were hard-coded systems that did not employ training or learning methods that are present in artificial intelligence based approaches. Furthermore, they lacked systematic performance assessment studies and therefore could not be objectively proved as accurate or reliable methods.

AI and expert system based models are known to be used in estimating the risks for various types of medical conditions in literature with an increasing popularity in the recent years [127]. Prediction based systems can be implemented to estimate more general medical parameters such as mortality rate or discharge time of the patients as well. Rajkomar et. al. constructed a predictive model using deep learning models to analyze the electronic health records (EHR) of patients and predict different outcomes including in-hospital mortality, prolonged length of stay and diagnosis by the time of discharge [148]. The proposed model was shown to be able to perform accurately on the 114 003 patient dataset formed from multiple healthcare centers without site-specific data harmonization. Deep learning architectures such as long short-term memory (LSTM) and time-aware artificial neural networks (TANN) were employed to implement the multi-output system. Area under the receiver operating characteristic curve (AUC) was chosen as the performance metric in the study. According to the results obtained by the authors, the proposed model was able to outperform traditional, clinically-used predictive models in all cases. The AUC values were reported to be 0,93; 0,85 and 0,90 (frequency weighted) for the model's performance in predicting in-hospital mortality, prolonged length of stay and diagnosis by the time of discharge, respectively.

Johnson et. al. proposed a machine learning based method to estimate the risks of 3 different events in patients with respect to Coronary computer tomography (CT) angiography image analyzes [149]. In their 2019 study, the authors formed a dataset from the coronary CT angiography results of 7350 patients. The dataset did not contain the images themselves, but

instead the interpretation of them by a medical expert, resulting in 64 score-based features for each patient. This dataset is then used to train four different machine learning algorithms: logistic regression, k-nearest neighbors (kNN), bagged trees, and simple ANN. These algorithms are used to predict the risks of all-cause death, coronary heart disease death, and nonfatal myocardial infarctions for each patient. The authors analyzed the results of the algorithms and concluded that kNN and bagged trees performed superior to the others with their performance criteria being AUC values. Both algorithms achieved an AUC of 0,77 in predicting all-cause deaths while kNN was slightly more successful in predicting coronary heart disease death and nonfatal myocardial infarctions with an AUC of 0,85 for both outcomes. The authors assumed a scenario involving the treatment procedure to prevent the predicted outcomes in order to compare their proposed method to a statistical method, namely Coronary Artery Disease Reporting and Data System (CAD-RADS) score. The scenario is visualized by adding prevalence information to receiver operating characteristic (ROC) information to obtain the modified ROC curves that contain the false-positive findings vs true-positive findings for kNN and CAD-RADS score predictions. The all-cause death and coronary heart disease death scenarios were analyzed at the points where the difference in treatment rates of proposed and statistical methods were largest. These points corresponded to false-positive/true-positive ratios of 8 and 45 for all-cause death and coronary heart disease death cases, respectively. Nonfatal myocardial infarction case was omitted due to low rate of follow-up returns from the patients. According to these results, the proposed algorithm was able to detect 22% and 24% more patients that would have an outcome of all-cause death or coronary heart disease death, respectively when compared to the CAD-RADS score results. Therefore, the machine learning based model proposed in this study was shown to improve the interpretation of vessel features to discriminate the patients who are more likely to suffer a negative outcome when compared to statistics based approaches.

Tomašev et al. proposed a method for prediction of acute kidney injury (AKI) using deep learning in their 2019 study [150]. Using the EHR of patients, the proposed system aimed to predict a future kidney deterioration alongside predicted future trajectories for blood tests. The conventional method of detecting AKI was reported to be measuring serum creatinine. However, since the abnormal levels serum creatinine were generally observed after renal injury, the predictive power of conventional methods was limited. The proposed system in their study was an RNN-based model operating sequentially to build a local workspace to

keep the track of relevant data. With each new admission of a given patient, the model provided updated estimate of the future risk of AKI alongside a measure of uncertainty of the estimate. The model was trained and tested employing a large, multisite retrospective dataset of 703 782 patients' EHR. The model predicted the probability of an AKI event taking place within 48 hours for the patients. The probability of AKI is determined with a regression model, which was transformed to a binary classifier that considers the regression results cumulatively, making a "AKI positive" prediction after some threshold. The reported accuracy of the system in terms of predicting any AKI event within 48 hours were apparently low at 55,8%. However, it should be kept in mind that earlier methods that were considered as benchmark prior to this study were able to predict the AKI episodes within 48 hours with accuracies no higher than 36% . It is highlighted in the study that early detection of the cases with a high risk of future AKI is critical since it would create the possibility of taking preventive action. The authors further discussed that providing the uncertainty associated with each prediction might help the medical professionals to discriminate the ambiguous cases from more clear ones and decide on the procedure to be followed accordingly.

There has been a number of different studies about stroke-related prediction and risk estimation problems in literature. One of the earliest applications of expert systems in stroke type prediction was the 1989 study carried out by Spitzer et. al [151]. In their study, an expert system, namely Microstroke, was proposed to perform a diagnosis according to the input provided by the user in a questionnaire format. Patient data from three stroke registries were used to form and train a modified Bayesian inference model that comprised the Microstroke's inference engine. The number of patients or features using the training were not provided, however the number of features can be inferred from the authors' statement of the questionnaire consisting of 38 items. The questionnaire, which was the method of providing input to Microstroke, expected the user to reply a number of yes/no and multiple choice questions to determine the risks of stroke related events, referred to as stroke types. The output parameters were given thrombosis, embolus, intracerebral hemorrhage (ICH), and subarachnoid hemorrhage (SAH). The proposed method was tested with 250 patients registered at Hamburg Stroke Databank prospectively. The performance measure of the study was chosen as the stroke type classification accuracy and the overall accuracy of the system was reported to be 72,8%. The accuracy of the proposed model in predicting individual stroke types are provided in Table 2.2.

Table 2.2. Accuracy of Microstroke for different stroke types

Stroke Type	Accuracy
SAH	91,7%
ICH	63,0%
Embolic	75,0%
Thrombolic	70,7%

One of the weaknesses of the study was the partial unavailability of some necessary information regarding the training data and model formation. Nevertheless, the study is believed to be significant since an acceptable accuracy in predicting stroke types was shown to be achieved using a simple, questionnaire-based input model that was realized using machine learning in 1989.

Estimating the post-stroke outcomes in patients is also among the most widely studied tasks in this field. Siegel et. al. investigated the effect of post-stroke distributed brain network disruption on impairment in multiple behavioral domains using machine learning models [152]. In their 2015 study, the authors considered resting functional connectivity (FC) and lesion topography alongside a number of behavioral capabilities namely, attention, visual memory, verbal memory, language and motor. The dataset used in the study was prospectively obtained from 132 stroke patients. The lesion topography was obtained using neuroimaging and manual segmentation while resting FC was estimated from the images via region of interest (ROI) based functional connectivity estimation. These two groups of data were used to form the structural data separately which were used as inputs in the proposed model. The neuropsychological assessment of behavioral capabilities was carried out systematically with a series of behavioral tests that included several evaluations of motor, language, attention, memory, and visual functions. The results of these tests were then brought together to form the behavioral data, which is regarded as the output or response parameters. In order to exploit the shared features between models, multitask learning was employed, which was based on the assumption that there are some shared features in brain functions regarding different domains or functions. L_1 regularization was chosen to prevent any bias towards the usage of shared or domain-specific weights during learning process. The objective function of multitask learning used in the study is given in Eq. 2.4.

$$\arg\min_{\omega} \lambda||\overrightarrow{\omega_0}||_2 + \sum_{k=0}^{K}\sum_{i=1}^{n}\left(\left(\overrightarrow{\omega_0+\omega_k}\right)^T\vec{x}_i - \vec{y}_i\right)^2 + \lambda||\overrightarrow{\omega_k}||_2 \tag{2.4}$$

where $\vec{x}$ and $\vec{y}$ are the FC and behavioral score vectors for patients i=1:n and λ is the regularization coefficient. ω_0 and ω_k correspond to domain general and domain specific weights, respectively. The final model optimized according to Eq. 2.4 was able to explain 28.7% of the variance across all patients and all five domains. The effect of FC and lesion deficits on behavioral domains were analyzed in a comparative manner. As a result of these analyses, it was asserted that impairment in visual and verbal memory as well as attention were more related to FC than lesion deficits while motor impairment had a stronger relation with lesion deficits. Language, on the other hand turned out to be similarly related to both inputs. These relations were explained in terms of prediction accuracies of the model, with FC and lesion deficits considered separately which can be found in Table 2.3.

Table 2.3. Prediction accuracies of FC and lesion deficit based models

Domain	FC Deficit	Lesion Deficit
Visual	36,4%	10,9%
Verbal	41,6%	18,7%
Attention	45,0%	32,3%
Motor	23,4%	44,8%
Language	51,1%	64,6%

Park et. al. proposed a Bayesian network model to predict post-stroke outcomes in their 2018 study [153]. Considering a more general set of events when compared to Tomašev et. al. [150], the authors constructed an inference engine to predict the 3 month functional independence and 1 year mortality of stroke patients. The dataset used in the study was formed from the data of 3605 patients registered in the Yonsei stroke registry, diagnosed with either ischemic or transient ischemic attacks. The raw dataset contained 76 risk factors. Later, feature selection was performed on the raw dataset and a number of datasets having different number of features were constructed. A Bayesian network model with its acyclic graph described in Eq. 2.5 was created to be trained with these datasets.

$$P(V) = \prod_V P(V_i|\pi(V_i)) \tag{2.5}$$

where P is the probability distribution associated with the graph, V are the random variables representing features and $\pi(V_i)$ is the set of parent nodes of i^{th} random variable V_i. Training of the Bayesian network model corresponded to finding the optimal model for P that represents the dataset most accurately with a hill-climbing search. The measure of this

representation accuracy was selected to be maximum description length (MDL) score. Two different feature selection methods were considered to perform dimensionality reduction in the dataset. The first method was filtering the dataset by information gain that is based on a performance measure, entropy for this case. The second feature selection method was reported to be a wrapper method, which involved training of the network exhaustively for each subset of features to find the optimal one. In order to determine the optimal model, the performance criteria was chosen as AUC. After a number of trials involving different feature sets suggested by the two feature selection methods, the most successful models were determined separately for predicting 3 month functional independence and 1 year mortality outcomes. The highest AUC obtained for 3 month functional independence was 0,889 and belonged to the 19 feature model formed by wrapper method. The AUC values for feature selection by information gain and base (no feature selection) cases remained at 0,875. A higher AUC obtained by the 19 feature model suggested that there might be some bias in the original dataset, resulting in a reduced performance when all 76 features are considered. For predicting 1 year mortality, a similar approach was followed and it was observed that the model with 24 features obtained by information gain feature selection method turned out to be slightly more successful. The AUC reported for this model was 0,895 while wrapper method and base case remained at 0,893 and 0,892; respectively.

The authors also discussed the most relevant predictors of post-stroke outcomes by analyzing the datasets formed by the two feature selection methods. It is suggested that D-Dimer was the strongest feature or risk factor that is associated with predicting the 1 year mortality in stroke patients. Initial National Institute of Health Stroke Scale (NIHSS) score and age also had a strong association for both feature selection based models. For predicting 3 month functional independence, the strongest feature was determined to be NIHSS score for both feature selection models. However, they suggested different features for the 2nd and 3rd strongest predictors. For the information gain model D-Dimer and CRP were the followers of NIHSS score while Age and D-Dimer were suggested to be 2nd and 3rd most effective features by the wrapper model, respectively. To sum up, the proposed method was not only successful in predicting the post-stroke outcomes, but also identified the most relevant risk factors associated with these outcomes.

Estimating the risk of a medical condition is different than detecting an already present condition. Therefore, it is important to realize the difference in diagnosis methods and risk

estimation methods for a given medical condition, stroke for this study. Although the objectives are different, it is believed that there is a stronger association between stroke outcome predictors and stroke risk estimators as both models aim to predict future events. That being said, a surprisingly low number of studies in the literature employed AI based methods to predict a form of stroke risk, two of which are presented in this section.

Khosla et. al. proposed an integrated machine learning approach to predict the risk of stroke in their 2010 study [154]. The proposed model included a novel automatic feature selection algorithm that picked the most useful features according to the conservative mean heuristic. Stroke was handled as the only parameter to be estimated in the study resulting a number of single output systems characterized by approaches taken during data imputation, feature selection and prediction steps. The dataset used in the study was obtained from a cohort of 5201 patients with 796 features. The authors discussed that the presence of such a large feature set made it inevitable that a large number of entries were missing. Therefore, the patients with more than 60% of missing entries were removed from the dataset, resulting in a 4988-patient set. Since most of the machine learning models, including the ones used in the study, were unable to work with incomplete datasets four different methods to handle the missing data were proposed:

- Column mean
- Column median
- Linear regression
- Regularized Expectation Maximization (EM)

For all four data imputation methods, the continuous variables such as blood sugar were imputed as it is while discrete-valued parameters such as presence of atrial fibrillation were rounded to the nearest discrete value. The authors tested the accuracy of each imputation method by segregating the dataset as training and test sets with a ratio of 9:1 and then estimated the missing values in the test set using the corresponding method. As a result of these tests, it turned out that linear regression was able to estimate the missing values with highest accuracy. However, column median was chosen to be adopted in the final model since it provided highest AUC values obtained in stroke prediction task, which was the main aim of the study.

Due to the high number of initial features, the authors proposed a novel feature selection algorithm that was inspired from the K-fold cross validation algorithm. Variance across different folds was taken into consideration to form the conservative mean parameter that is expressed as the "mean minus standard deviation" of the AUC of a regression model. By this way, it was made possible to evaluate the effect of each feature on prediction performance more clearly. The authors tested this proposed method of feature selection against two other methods, namely forward feature selection and L_1 regularized linear regression. The results showed that the proposed conservative mean method was superior to both alternatives as it yielded higher AUC values for both prediction algorithms considered in the study.

The algorithms used to predict stroke events in the study were support vector machines (SVM) and margin based censored regression (MCR). These two algorithms were also compared to each other and it was observed that MCR outperformed SVM with respective AUC values of 0,777 and 0,774. The authors further compared the results of the proposed model to a number of different methods in literature in two levels. The feature selection method proposed in the study was compared to the method of Lumley et. al. by using the two prediction algorithms on their 16-feature dataset. Results showed that the proposed feature selection method yielded higher AUC values for both prediction algorithms. The prediction algorithms were also compared to a statistical prediction method, namely the Cox proportional hazards model in terms of AUC and concordance index. The authors suggested that both MCR and SVM models were able to achieve better AUC and concordance index results than Cox model employing the identical feature selection methods. The results of these comparisons are provided in Tables 2.4 to 2.6.

Table 2.4. AUC of Lumley's feature selection method compared to the proposed method

Feature Selection Method	SVM	MCR
Conservative Mean	0,774	0,777
16 Features (Lumley)	0,753	0,765

Table 2.5. AUC of Cox models compared to the different models used in the study

Model (Predictor + Feature Selection Method)	AUC
MCR + Conservative Mean	0,777
SVM + Conservative Mean	0,774
SVM + 16 Features (Lumley)	0,753
SVM + Forward Selection	0,751
Cox + L_1 Regularized Linear Regression	0,747
Cox + 16 Features (Lumley)	0,734

Table 2.6. Concordance index of Cox models compared to the different models used in the study

Model (Predictor + Feature Selection Method)	Concordance Index
MCR + Conservative Mean	0,770
MCR + 16 Features (Lumley)	0,757
SVM + Conservative Mean	0,760
SVM + 16 Features (Lumley)	0,747
Cox + Conservative Mean	0,737
Cox + 16 Features (Lumley)	0,730

Weng et. al. performed a prospective study to detect and label high-risk individuals by means of cardiovascular diseases, including stroke in their 2017 study [7]. Four different machine learning algorithms were proposed and compared with each other, alongside a conventional, guideline-based algorithm. Although an explicit definition of CVD was not provided, the diagnosis codes of UK National Health Service (NHS) that were considered as CVD in the study were stated to be considered. According to the code groups stated in the study, the authors considered a number of coronary (ischemic) heart conditions and cerebrovascular conditions, including stroke, as CVD. In parallel with this approach, the models developed in the study aimed to predict whether a patient will have CVD in 10 years as their sole output parameter. The dataset used in the study was formed prospectively from the Clinical Practice Research Datalink (CPRD) that anonymized EHRs from 378 256 patients that were healthy by the time of initial acquisition. 30 different features were considered for each patient in the dataset, 8 of which comprised the risk factors considered in 2013 ACC/AHA guidelines. The dataset was then used to train and validate four different machine learning algorithms given as logistic regression, random forest, gradient boosting machines and a 3 layer ANN structure denoted as simply “neural network”. The machine learning models were trained and used individually with 2-fold cross validation against overfitting. Maximization of AUC

was chosen to be the performance metric for the study. The baseline model following the 2013 ACC/AHA guidelines to predict CVD in patients was tested using its 8-feature input set as a reference point. The accuracy and AUC values obtained by this base case was given as 62,7% and 0,728, respectively. The four models proposed in the study were tested using the 30-feature input set and the AUC values given in Table 2.7 were obtained.

Table 2.7. AUC values of methods proposed in the study

Method	AUC
Random Forest	0,745
Logistic Regression	0,760
Gradient Boosting Machines	0,761
Neural Network	0,764

As can be seen from Table 2.7, The neural network model performed better than other models in the study and all machine learning models obtained higher AUC values than the baseline model. The authors also highlighted the flexibility of machine learning algorithms by means of the features to be considered, contrary to fixed-feature characteristics of statistical prediction models. The significance of each feature or risk factor was also discussed according to the results of model-specific analyzes carried out for machine learning models. The features with highest contribution to the prediction process were named as top-ranking risk factors. As a result of these analyzes, it is observed that age, gender and smoking were among the top-ranking risk factors that were also considered by ACC/AHA guidelines [147]. There were, however, other significant risk factors such as severe mental illness, oral corticosteroid usage or triglyceride levels that were not found in the ACC/AHA model. Diabetes on the other hand, was the only risk factor that was not considered among the top-ranking risk factors in any of the machine learning algorithms despite being prominent in the ACC/AHA model. 3 most important risk factors determined by each machine learning model, as well as ACC/AHA (stratified by gender) is provided in Table 2.8. It is observed that only social feature in Top 3 was Townsend Deprivation Index (TDI), ranking 3rd in the logistic regression model.

Table 2.8. Top 3 risk factors for models used in the study

ACC/AHA (Men)	ACC/AHA (Women)	Logistic Regression	Random Forest	Gradient Boosting Machines	Neural Network
Age	Age	Ethnicity	Age	Age	Atrial Fibrillation
Total Cholesterol	HDL Cholesterol	Age	Gender	Gender	Ethnicity
HDL Cholesterol	Total Cholesterol	TDI	Ethnicity	Ethnicity	Oral Corticosteroid

Although there are many examples of AI based models in prediction of different diseases, their usage in predicting stroke is mostly limited to outcome prediction of an already occurred stroke event [152,153]. The few stroke risk estimation studies in literature, despite proving superior to conventional approaches by means of prediction performance and risk factor identification, had a major drawback in terms of output parameters. These models either regarded stroke as a single parameter or included it alongside different medical conditions which were also merged into a single parameter to be predicted. It should, however, be kept in mind that ischemic and hemorrhagic stroke events exhibit some distinct characteristics due to their different mechanisms. As discussed in medical background section, there are some important factors that affect the ischemic and hemorrhagic stroke risks in different directions. For this reason, it is believed that different types of stroke should be handled differently in the prediction task.

Even though a number of different and mostly diverse algorithms were employed and tested in each of the stroke risk estimation studies presented above, the general approach was in the form of “choose the best, discard the rest”. On the other hand, ensemble learning could have been employed to benefit from having a number of diverse predictors in hand and therefore obtain better results as suggested by Kuncheva and Whitaker [155].

Last but not least, it is observed that the performance of proposed methods in literature, usually provided in terms of AUC, were limited as no model achieved an AUC of 0,8 or greater. Therefore, in this study it is aimed to overcome these shortcomings by proposing a novel, two stage model that employs ensemble methods to accurately predict different types of stroke with a large set of risk factors.

3. DATASET

In this study, a novel dataset containing medical information of patients is formed to be used in the training and testing stages of the proposed stroke risk estimator. In order that the risk estimator makes accurate predictions, the dataset used for its training must represent the sample space as accurately as possible. For the case of this study, the sample space represents the general population and its stroke related parameters. For this reason, any feature of medical significance related to stroke is aimed to be included in the dataset to provide an accurate representation of the patients' medical condition.

An extensive search of a suitable dataset among readily available ones were carried out initially. However, the available datasets were considered as unsuitable primarily because they deemed insufficient in terms of either sample size or number of features. There were some datasets with sufficiently large number of samples and features such as International Stroke Trial (IST) dataset with 19 435 patients and 112 features [156]. However, the features of IST were largely focused on anticoagulants and lacked the diversity that is necessary to consider stroke risk on a wider scope. Few databases that are believed to be suitable to be used in the study such as Swedish Stroke Registry (Riksstroke) were not available for public usage [157]. For these reasons, the dataset to be used in this study is determined to be created from scratch, taking advantage of the freedom in determining certain properties of the dataset such as the features to be included or the number of patients to be considered.

3.1. Determination of Raw Data

In order to estimate the stroke risk in patients, the risk factors, mostly in the form of medical information must be considered in the dataset. However, such medical information can only be found in health records of patients which are not stored in a dataset format under normal circumstances. For this reason, in order to form a proper dataset, the medical information was first needed to be obtained in a raw form and then processed into features.

Since it is aimed to consider every feature that could possibly be related to stroke risk in patients, the raw data elements to be acquired is determined accordingly. With the aid of a medical professional (co-advisor of this study) a list of different types of health records that

could contain information related to stroke and its risk factors were prepared. This list formed the basis for the raw data used in the study. Health records that were included in the raw data list alongside some example features related to each record are provided in Table 3.1.

Table 3.1. List of health records included in raw data and example features

Category	Examples
General Information	Age, Sex, Height & Weight
Anamnesis and Epicrisis	Main problem of the patient, recent childbirth, surgical history
Family History	Stroke or cardiovascular diseases in family
Diagnosis	Diabetes, Hyperlipidemia, Hypertension
Behavioral Factors	Alcohol consumption & smoking
Hemogram	Hematocrit, Hemoglobin, MCV (Mean Corpuscular Volume)
Biochemistry	Urea, Cholesterol, Triglyceride, AST, ALT, Minerals
Hormones	TSH (Thyroid Stimulating Hormone), Free T3 & T4
Complete Urine Test	Blood, Glucose & Ketone presence in urine
Medical Imaging Reports (Cerebral and Cardiovascular System)	CT (Computerized Tomography), MRI (Magnetic Resonance Imaging), Angiography, Echocardiography

3.2. Acquisition of Raw Data

The list of health records provided in Table 3.1 was provided in the application made to Gazi University Clinical Research Ethics Committee regarding the acquisition of raw data for the study. The study and data acquisition were approved by the committee's decision on 09.12.2019 with decision number 24074710-604.01.01-18, a copy of which can be viewed in Appendix-1. Following the approval of the committee, the raw data was requested from Gazi University Faculty of Medicine. With the approval of chief medical office, the requested data is acquired anonymously and retrospectively. In this context, no personal information regarding the patients such as name, patient number or contact information was accessed or used by the researcher. The raw data consisted of the aforementioned medical information of patients that were admitted to Gazi University Hospital inpatient and outpatient clinics between January 2007 and October 2019. Total number of patients included in the study is 3000. The discrimination in the dataset was made according to whether the patients had a stroke event that can be verified by the available information or not. The inclusion criterion for the Test (Positive, Stroke+) group was having at least one recording of ischemic stroke, hemorrhagic stroke, TIA or unidentified (general) stroke. The

only exclusion criteria apart from insufficient data was the AVM and aneurysm related events as explained in Section 2.1.2. The distribution of patients by means of gender and stroke record was even, as shown in Table 3.2.

Table 3.2. Distribution of patients in the dataset with respect to gender and stroke record

Number of Patients	Male	Female
Positive Group (Stroke+)	750	750
Negative Group (Stroke-)	750	750

The test group consisted of patients that have had a stroke event between 2007-2019 with multiple stroke cases included. For these patients, the health records dating 2-12 months prior to the first record of stroke within the given time period were acquired as raw data. The large variance of health records' intervals prior to records of stroke made it necessary to adopt this 10 month-wide margin in order to keep the number of samples high enough. For the control group patients, the data acquired is ensured to span at least 12 months, ensuring that these patients did not have a reported stroke event within 12 months from the date of admission. For this reason, patients registered after October 2018 were not included in control group, even if they did not have a stroke diagnosis. Taking these facts into account, the output of the proposed model can be interpreted as the prediction of whether the given patient is expected to have a stroke event in 2-12 months from the time of admission.

The patients were between 7 and 94 years old, with median age being 61.6 years old. The four output parameters determined to be used in the study are ischemic stroke, hemorrhagic stroke, TIA and general stroke. Here, general stroke is the output parameter that represents whether the given patient have had a stroke or not, regardless of its type. Therefore, it includes both unidentified strokes and other three types of strokes. The distribution of patients with respect to the output parameters and gender are given in Table 3.3.

Table 3.3. Distribution of patients in the dataset with respect to gender and stroke type

Number of Patients	Ischemic	Hemorrhagic	TIA	Stroke
Male	541	153	155	750
Female	499	192	91	750

It should be noted that the first 3 columns of each gender do not add up to the 4th column because of the patients with multiple records for different types of stroke.

3.3. Dataset Formation

After the acquisition of raw data, the features to be included in the dataset are determined considering the available information and possible risk factors according to the medical literature, as well as the expertise of a medical expert (thesis co-advisor). The information that is known to be unrelated to stroke risk is largely discarded. However, some features with no proven relation to stroke risk are still considered due to the assumption that there are still unknown stroke risk factors as discussed in Section 2.1.4. The features that were absent in more than 60% of the patients were discarded to preserve reliability. Among these the only features with a known relation to stroke risk were height and weight of the patients, which were used to calculate the body mass index [158]. Instead, a closely related feature, obesity, was included in the feature set [159].

The formation of dataset from the raw data is carried out manually. The age information, measurements such as SBP and laboratory results such as hemogram or biochemistry in numerical form were taken as they are and treated as continuous variables. The remaining data, which constituted approximately 50% of the feature set was processed by reading and evaluating the records individually. The record types used in the formation of these features are as follows:

- Anamnesis (Including sex, behavioral factors and family history)
- Epicrisis
- Diagnosis
- Medical Imaging Reports

The information contained in these records were in the form of verbal expressions. Therefore, these expressions are converted into features with discrete and binary values. For instance, a patient with the sentence “Had a myocard infarct” in his/her anamnesis was transformed to a “1” under the feature “Myocard Infarct” while the expression “2nd degree mitral insufficiency” was converted to a “2” under the feature “Mitral Insufficiency”.

The patients with missing critical information, namely, stroke, age and sex were discarded from the dataset. Furthermore, the patients with more than 60 missing features were also discarded. The resulting dataset contained 3000 samples with 124 input and 4 output features.

3.4. Input Features

The dataset used in this study consists of 124 input features which are provided with their brief descriptions in this section. The output features were already provided in the preceding section; hence they will not be addressed again.

Age: Patient age at the time of first diagnosis.
Age Group: The discretized version of patient age at the time of first diagnosis. 7 groups are defined, proportional to the age.
Alanine Aminotransferase (ALT SGPT): An enzyme synthesized by liver cells that plays an essential role in metabolic activities.
Alcohol: Alcohol intake in terms of drinks per day. 10 g of pure alcohol is considered as one drink according to International Alliance for Responsible Drinking [160]. Treated as a discrete variable.
Anticoagulant: History of anticoagulant usage such as Coumadin or Coraspin.
Anxiety: Anxiety disorders including generalized anxiety disorder, panic disorder, and phobia-related disorders. Occasional anxiety is not considered in this context.
Aorta Aneurysm: Expansion of a section in Aorta wall.
Aortic Insufficiency: Degree of aortic insufficiency or regurgitation that is the leaking of the Aorta. Treated as a discrete variable.
apTT: Activated Partial Thromboplastin Time. Considered as a measure of coagulation of the blood.
Arrhythmia: Irregular heartbeat
Aspartate Transaminase (AST Sgot): An enzyme that is generally measured to detect any abnormalities in the liver.
Atherosclerotic heart disease (ASKH): Cardiovascular diseases due to atherosclerosis
Atrial Fibrillation and Flutter: Presence of any of the two Tachycardia types with sinus pattern disruption.

Blood in Urine: Presence of blood in urine or hematuria either in gross or microscopic form.
Blood Urea Nitrogen (BUN): Amount of urea nitrogen present in blood.
Basophil: A type of white blood cell. They are the largest type of granulocyte.
BNP (B type Natriuretic Peptide): A neurohormone that is considered as a marker of congestive heart failure.
Bradycardia: A record of resting heart rate of under 60 beats per minute.
Bypass: History of coronary artery bypass surgery.

Cancer: Present or previous records of cancer in patient.
Coronary Artery Disease: Reduction in blood flow to the heart tissue by plaque buildup in the wall of the related arteries.
CRP (Immunoturbidimetric): A quantitative test to determine the amount of C-Reactive Protein in human serum.
Cardiomyopathy: Chronic illness of heart muscle.
Carotid Artery Stent: History of carotid artery stent placement surgery.
Carotid Plaques: Presence of plaques in carotid artery
Creatine Kinase (CK): An enzyme that catalyzes the conversion of phosphocreatine which is the energy reservoir for regeneration of ATP in body.
Creatine Kinase MB (CK-MB): Myocardial fraction of creatine kinase.
Cerebrovascular Disease: Diagnosis of any cerebrovascular disease, excluding stroke events.

DBP: Diastolic Blood Pressure that is the minimum blood pressure between two heartbeats, expressed in mmHg
Density: Density of the urine.
Diabetes: Diabetes mellitus, either type 1 or type 2.
Dysphasia and Aphasia: Dysphasia is a condition that affects the verbal ability of the patient. Aphasia is the more severe form of dysphasia, resulting in complete loss of speech and comprehension of spoken or written language.
Dyslipidemia: Abnormality of amount of lipids in the blood, either in the form of hyperlipidemia or hypolipidemia.
Dyspnea: Shortness of breath, temporary dyspnea after heavy exercise is not included.
Unidentified Disorders of the Circulatory System: Any unidentified disorder observed in the circulatory system such as hypotension or gangrene.

DVT: Deep vein thrombosis, thrombus formation in one or more of the deep veins in the body, mostly in legs.

Eosinophil: A variety of white blood cells that are responsible of fighting infections and parasites in humans.
Epilepsy: A neurological disorder characterized by recurrent seizures involving abnormal brain activity due to sudden electrical discharges in brain.
Erythrocyte: Count of Erythrocytes in unit volume of blood.

Family Diabetes: Presence of diabetes in the patient's family members.
Family Hypertension: Presence of hypertension in the patient's family members.
Family Stroke: History of any stroke event in the patient's family members.
Family CVD or MI: Presence of cardiovascular disease or myocard infarct history in the patient's family members.
Ferrum: Iron level in blood.
Fetal PCA anomaly: An anomaly of the posterior cerebral artery (PCA) that supply oxygenated blood to the occipital lobe of the brain. It is characterized by the posterior communicating artery (PCOM) being larger than the P1 segment of the PCA.
Fever: Persistently high body temperature, above 37.5 degrees Celsius.
Fainting: History of any syncope event in the patient.
Free T3: Amount of free T3 hormone in blood, considered as one of the most important thyroid hormones.
Free T4: Amount of free T4 hormone in blood, considered as one of the most important thyroid hormones.

Gamma-Glutamyl Transferase (GGT): An enzyme that catalyzes the transfer of gamma-glutamyl functional groups, mostly found in liver.
Glycosylated Hemoglobin (HB A1c): The amount of glycosylated hemoglobin in the blood, used to assess the status of glucose metabolism in body.
Glucose (Fasting): Glucose level in blood, measured after 8+ hours of fasting.
Glucose in Urine: Presence of glucose in urine above 0.8 mmol/L.
Goiter: Abnormal enlargement of the thyroid gland.
Gender: Sex of the patient, a binary variable with 1 corresponding to female.

GFR: Glomerular Filtration Rate, an important indicator of kidney function. Calculated according to Ckd-Epi formula and expressed in ml/1,73min x m^2 .

HCO3 (Bicarbonate): A vital component for the acid-base homeostasis in the human body.
HDL Cholesterol: High-density lipoprotein, often referred to as "good" cholesterol.
Heart Disease: Other illnesses related to heart that are not coronary artery disease.
Hematocrit: The volume percentage of red blood cells in blood.
Hemoglobin: Oxygen-transport metalloprotein in the red blood cells that contains iron.
Hyperlipidemia: Abnormally high levels of any or all lipids or lipoproteins in the blood.
Hypertension: Persistently elevated blood tension with SBP and DPB readings above 140 and 90 mmHg, respectively.
Hypothyroidism: Underactivity of thyroid characterized by its unability to synthesize enough thyroid hormones.

ICA Stenosis: Stenosis or occlusion of Intracranial Artery. Treated as a discrete variable that takes the value 1 for stenosis and 2 for occlusion.
Iron-binding Capacity: The blood's capacity to bind iron with transferrin. Its test also indirectly measures transferrin level as it is the main carrier of iron in blood.
Iron deficiency anemia: Anemia caused by lack of iron in blood. The accompanying transferrin levels are generally high while serum iron is below normal levels.
Infective Endocarditis: Infection of the inner surface of the heart, usually the valves.
INR: Coagulation duration of blood in terms of International Normalized Ratio.
Interstitial Lung Disease: A disorder that causes progressive scarring of lung tissue reducing the effective lung capacity.

Ketone in Urine: Presence of ketones in urine, considered as an indicator of poorly controlled diabetes and dietary disorders.

LDL Cholesterol: Low-density lipoprotein, often called the “bad” cholesterol.
Lymphocytes (Percent): Percentage of Lymphocytes to the total number of white blood cells in blood. Can be used as a marker of viral infections and lymphoma.
Leukocytes: Total leucocyte (white blood cell) count per mm^3 of blood.

MCH: Mean Corpuscular Hemoglobin, refers to the average quantity of hemoglobin present in a single red blood cell.

MCHC: Mean Corpuscular Hemoglobin Concentration, corresponds to the average concentration of hemoglobin in the red blood cells.

MCV: Mean Corpuscular Volume, represents the average volume of red blood cells in blood.

Migraine: Headache disorder that is characterized by recurrent headaches that are moderate to severe.

Mitral insufficiency: Degree of mitral insufficiency in the patient, treated as a discrete variable.

Monocyte: A phagocytic blood cell that is capable of destroying invading bacteria or other foreign material. Expressed as a percentage of leukocytes.

MPV: Mean platelet volume, is the measurement of the average size of platelets found in blood.

Murmur: Presence of abnormal heart sounds that could be a sign of heart defects.

Myocard Infarct: History of a MI, also known a heart attack.

Nitrite: Presence of nitrite in urine. Usually considered as a marker of urinary tract infection.

Neutrophil : A type of leukocytes that comprise up 70% of all white blood cells in humans.

Obesity: Present or recent record of body mass index (BMI) over 30 kg/m^2.

Pacemaker: Presence of a device that regulates the heartbeat

Palpitation: Perceived abnormalities of the heartbeat, usually in the form of strong and/or fast heartbeat. The patient's description was considered for this feature.

Pulse: Frequency of heartbeat, expressed in beats per minute (bpm).

PCT: Represents the amount of procalcitonin in blood. Generally used to detect presence of a bacterial infection or sepsis.

PDW: Platelet Distribution Width, is the measure of the variance of platelet sizes in blood. It is considered as an indicator of platelet function and activation.

pH: pH level of urine.

Pneumonia: An inflammatory disease of the lung primarily affecting the alveoli.

Potassium (K): The amount of potassium in blood which is an electrolyte that is essential for proper muscle and nerve function.

Pregnancy: History of pregnancy and/or childbirth within 1 year of diagnosis.

Prothrombin time: Measured blood clotting time, indicated the coagulation speed.
Protein: Presence of protein, especially albumin in urea. Generally considered as a marker of kidney diseases.
Pulmonary Embolism: Blockage of an artery in the lungs due to an embolus formed elsewhere in body.

RDW: Red cell Distribution Width, is a measure of the variance of the red blood cell size.

SBP: Systolic Blood Pressure that is the maximum blood pressure during one heartbeat, expressed in mmHg
Sedimentation: A measure of sedimentation amount of red blood cells in serum blood in 1 hour, expressed in terms of millimeters.
Shortness of breath: The reported feeling of not being able to breathe well enough.
Smoking: Current or past record of regular smoking, expressed in packs per day (ppd). Quitters' ppd values were reduced depending on the duration of smoking as well as quitting date whenever available.
Sinus Tachycardia: Record of elevated heart rate without the distortion of beating pattern. Cases with heart rate greater than 100 bpm at rest were included.
Sinus Thrombosis: Thrombosis due to blood clot within the Dural venous sinuses. Includes cavernous, lateral and superior sagittal sinus thrombosis cases.
Sleep Apnea: Record of pauses in breathing or abnormally shallow breathing during sleep.
Sodium (Na): Amount of sodium in the blood serum which is an essential mineral for muscle and nerve function.

Total Cholesterol: A measure of the total amount of cholesterol blood. It includes LDL, HDL, Very Low Density Lipoprotein (VLDL) and any other types of cholesterol
Transferrin saturation: The ratio of serum iron to the total iron-binding capacity of the available transferrin, expressed as a percentage.
Triglyceride: Corresponds to amount of triglyceride in blood, expressed in mg/dl. Considered as the main constituents of body fat as triglycerides are the form in which excess fat is stored in body.
Thrombocyte: Number of platelets in blood. Thrombocytes are one of the main components in the coagulation process of blood.

Troponin I: One of the three proteins that regulate the muscle contraction in skeletal and cardiac muscles. Generally considered as a marker of MI and acute coronary syndrome.
TSH: Thyroid Stimulating Hormone level in blood, used alongside free T3 and T4 levels to assess thyroid function.

Uric Acid: Concentration of uric acid in blood. Generally regarded as a marker of diabetes and gout.
Urobilinogen: Amount of urobilinogen in urine, which is a byproduct of bilirubin reduction in body. Considered as an indicator of liver disease or damage.

Valve Replacement: History of a heart valve replacement surgery.
Ventricular Salvo: Record of 3 or more consecutive Ventricular Premature Contractions (VPC), also known as ventricular triplets.
Vertebral Occlusion: History of occlusion or stenosis of vertebral artery. Treated as a discrete variable that becomes “1” for mild occlusion and “2” for severe occlusion or stenosis.
Vertigo: Persistent or severe vertigo.
Vitamin B12: Amount of Vitamin B12 in serum. Also known as cobalamin, Vitamin B12 is a water-soluble vitamin that is essential for the body for proper red blood cell formation, neurological function, and DNA synthesis.
Vitamin B12 Deficiency Anemia: Diagnosis of anemia due to the deficiency of Vitamin B12. Also known as pernicious anemia.
Vitamin D Deficiency: Deficiency of Vitamin D, determined from the serum concentration of calcifediol which is the main serum-solved version of Vitamin D.

3.5. Missing Data Imputation

Due to the fact that a large feature set is considered for the dataset used in the study, it was inevitable to have missing values of some features for almost all samples. The dataset used in this study was formed using the medical records of the patients that were admitted to various clinics of Gazi University Hospital. Since there is a wide range of different examinations, tests and analyses that are carried out depending on the patient characteristics, symptoms and the medical branch(es) applied to by patient, the available medical data varies considerably. Because of this fact, even after discarding patients with less than 60 features

present, the resulting dataset had inevitable missing data. A remedy to this missing data problem must be provided as the methods employed in this study required gapless datasets to function properly.

Discarding the samples with any missing data was not a viable option since more than 94% of samples contained at least one missing value. As discussed earlier, a threshold of 60 features were employed to filter the samples with too many missing information. In order to preserve the wide variety of the feature set, only a small number of features that were absent in most of the samples were discarded. Therefore, no further modification of the feature set was considered either.

As discussed by Khosla et. al. [154] there are a number of data imputation methods that can be used to handle the missing data. Among these, column median was deemed a suitable method to be applied in this study. That is, for each feature (column), the missing values for that feature is replaced by the median of the available values for that feature in the dataset. Data imputation with column median is preferred to that with column mean as it is believed to yield a more realistic dataset thanks to its immunity to extremely small or large values that may introduce an unnatural bias [161]. Furthermore, column median method was shown to yield higher classification performance than other methods, including linear regression and regularized EM with the dataset used by Khosla et. al. which was also a stroke-related dataset formed using medical information [154]. For these reasons, column median was chosen to be applied in this study as the missing data imputation method.

For continuous valued features, the imputed values were stored in double format which is the default numeric format of Matlab. For all of the discrete valued features, the median value turned out to be 0 and the missing values were imputed accordingly. This result can be considered to be consistent with the nature of the medical data since a diagnosis would normally not be recorded if it is not present in a patient and vice versa. In other words, if a diagnosis or record of a condition is not present for a patient, assuming that the patient does not have that condition is the general practice if further examination is not possible.

4. PROPOSED MODEL

In this study a novel, two stage model that can accurately estimate the risk of stroke events in patients is proposed. Its operation and performance are demonstrated using the dataset formed within the scope of the study.

The proposed model consists of a clustering module followed by a risk estimation module in a cascaded manner. In other words, the output of the clustering module is provided to the risk estimation module as its input. With the proposed model, two different machine learning approaches are used, namely unsupervised and supervised learning in clustering and risk estimation modules, respectively.

The first stage of the model performs clustering of the dataset which modifies it by adding the resulting clustering information for each sample. This modified dataset, now called clustered dataset, is then provided to the second stage of the model which includes the classifiers used to perform the stroke risk estimation task. The model is designed to adapt itself to the sample space comprised by the input dataset. Therefore, the structure of the risk estimation module exhibits dynamic behavior, adapting the contribution of the classifiers to be used on the output depending on their performance on the clustered dataset. This unique property maximizes the proposed model's invariance to different datasets by eliminating the effect of possible bias. The model uses the clustered dataset to estimate the general stroke risk first. After estimating the general stroke risk, the model may or may not use the information obtained from this stage to discriminate the patients with a very low stroke risk depending on the operation mode selected by the user. Finally, the estimation of three different types of stroke risks is carried out within this module.

Since the stroke information is expressed in binary format in the dataset, as well as the test data, which is simply a subset of the main dataset, in order to enable an objective evaluation of the performance, the final output must also be expressed in binary format. In this context, a "1" written under a stroke type in the dataset, corresponded to the given patient had a stroke of a given type and vice versa. The output of the proposed model was converted to binary format via a thresholding function in the last stage to make error calculation possible. By this way, it is made possible to test the proposed model with the dataset formed in this study.

As discussed in Section 3.2, the output of the model can be interpreted as the prediction of whether a given patient will have a stroke event of selected type within 1 year from his/her admission. The flowchart summarizing the general operation of the proposed model is given in Figure 4.1.

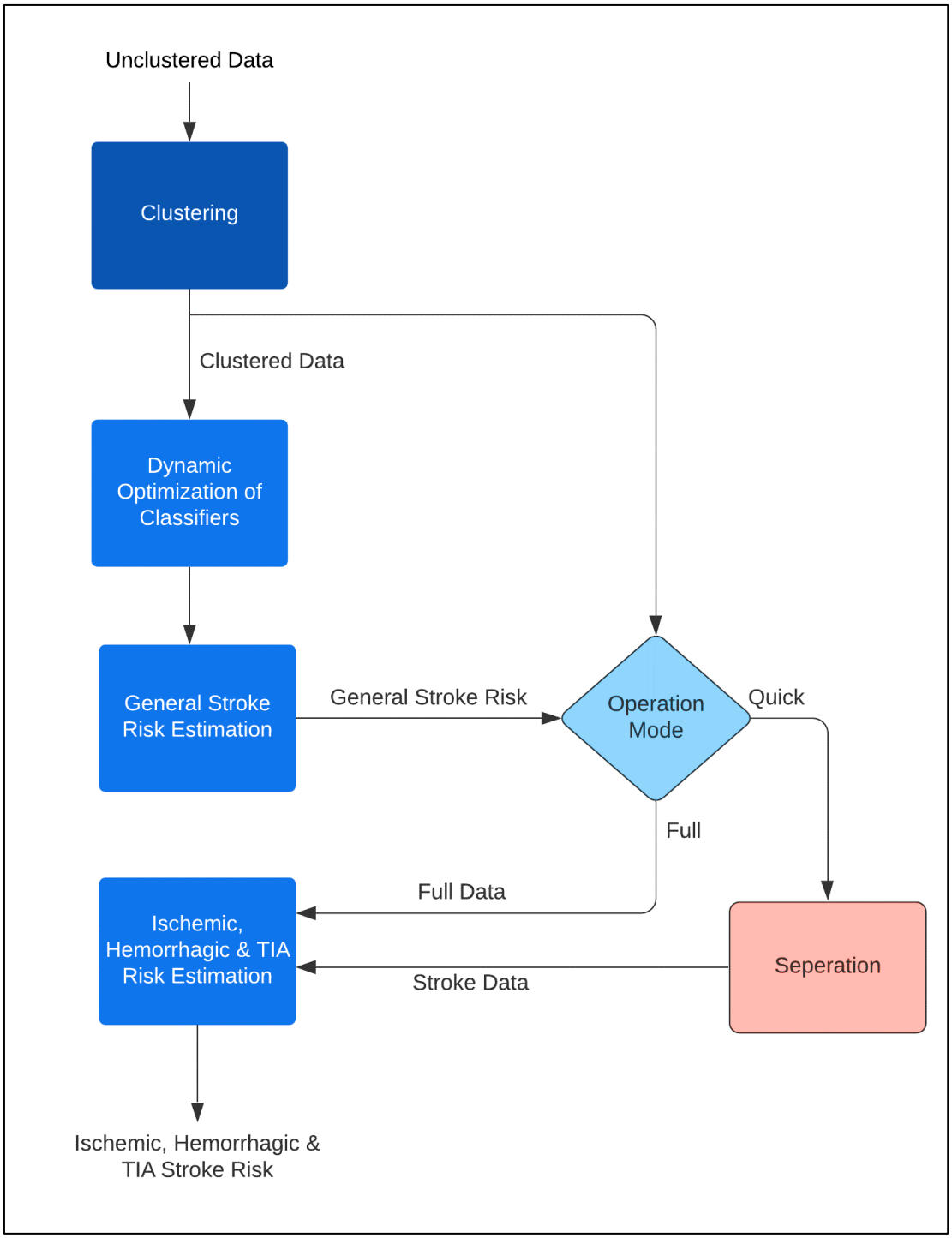


Figure 4.1. Flowchart summarizing the general operation of the proposed model

The clustering module, illustrated in dark blue in Figure 4.1, is employed based on the idea that grouping the patients with similar characteristics in terms of a general evaluation of their medical features could provide additional information for the stroke risk estimation task. In other words, it is suggested that clustering corresponds to an objective assessment and subsequent grouping of patients considering all of their features. This process can uncover some useful information that could then be used to improve the prediction accuracy of the stroke risk estimator module. The information uncovered by clustering process is actually an implicit evaluation of the patients based on their features, which can be treated as additional predictors of stroke. This idea is further discussed on Section 4.1 and its validation is provided in Section 5 by comparing the prediction performance of various models using only the original data to the models employing clustering results alongside the original data.

The risk estimation module consists of the building blocks given in medium blue in Figure 4.1. As its name suggests, this module ultimately aims to predict the risk of different types of stroke according to the patient data provided. In order to maximize the prediction performance, a dynamic variant of ensemble methods is proposed and implemented for this module.

Ensemble methods are a group of machine learning techniques that aim to improve the performance of the systems in performing tasks such as classification, regression or estimation by employing multiple models. The models could be trained and used on either a subset of the original dataset or the entire dataset, depending on the ensemble method used. The outputs of these models are then combined by either voting (typically in classification problems) or averaging (typically in regression problems). The contribution of each model in the final output can be equal in some applications while many contemporary ensemble methods employ weighted calculation of the final output [162]. In these methods, the weight of each model is generally assigned according to its performance on the training set. After the training stage is completed, these weights are held constant.

In this novel approach, ensembles of different classifier algorithms are formed and used for each stroke type similar to conventional ensemble methods [163]. However, unlike conventional ensemble methods, the weights of each classifier are not fixed after the initial training stage. Instead, they are dynamically adjusted at each run, depending on their performance on the validation part of the clustered dataset which is treated as the training

dataset for the ensemble weights. The adaptive nature of the clustering module enables generation of “new” training sets that can be used to optimize the weights at each run, without the necessity of re-training the models used in the ensemble individually. By this way, the classifiers do not lose their learned ability to make predictions towards a definitive task, stroke risk estimation for this study.

The data input to the ischemic, hemorrhagic and TIA risk estimation (IHTRE) block is determined by the operation mode block given in Figure 4.1 in light blue. This block has two possible output values, either of which can be selected by the user. If the “Full” mode is selected, the operation mode block will forward the clustered dataset to the IHTRE block without any modification. If the user selects “Quick” as the mode of operation, the result of general stroke risk estimation will be forwarded to the separation block (Illustrated in peach in Figure 4.1.) alongside the clustered data. In the separation block, the patients with very low risk of (general) stroke are detected and separated from the dataset which is then forwarded to the IHTRE block. By this way, the computational burden on IHTRE block is reduced by a factor proportional to the ratio of patients with low risk of stroke to total number of patients in the dataset. This mode inevitably sacrifices some prediction accuracy while reducing the computational cost since the separated patients are automatically assumed to have very low risks for all stroke types which may not always be the case.

4.1. Clustering Module

As discussed in Section 2.2.3, clustering is a machine learning method that employs unsupervised learning to separate the data into groups, often referred to as clusters. One of the main advantages of clustering is that it does not need labeled data. This is considered as a desired property in many applications including the ones that work with medical data as labeling a large dataset can be a tedious task. As a result of this, the output of a clustering algorithm does not contain any labels other than group or cluster-related information obtained in the clustering process. In some applications such as data mining, the main objective is to understand the structure of data [129]. Therefore, the output of a clustering algorithm can be studied to understand the characteristics of the samples in each cluster and label the clusters accordingly. In other cases, including this study, the main focus is not the information contained in the clusters itself, but instead how that information could be used towards a given objective, estimating the stroke risk accurately for this study.

In this study, the clustering module is used to provide an unlabeled categorization of patients according to their medical information, represented by 124 features in the dataset. Therefore, the clusters formed by this module are considered to represent an implicit yet impartial evaluation of their members. This implicit information, expressed in terms of clustering results, could then be used to provide additional guidance to the risk estimation module to ultimately improve the prediction performance. Here, considering the fact that the patients with similar evaluation are more likely end up in the same cluster, it is hypothesized that the evaluation of the patients conveyed by the clusters contain useful information regarding stroke risk. This hypothesis is practically proved in Section 5, by demonstrating that the inclusion of clustering results in the dataset provided to the risk estimation module actually improves the prediction performance regardless of the estimation method or stroke type.

The model proposed in this study aims to predict the risk of stroke in patients by considering their medical information. The idea that grouping or clustering the patients according to their medical characteristics should not be considered as exclusive to machine learning. In some medical practices, it could be useful for the doctor to divide patients into groups according to their medical information and apply group-specific diagnosis or treatment [164]. The groups formed this way could consider different aspects of available information and therefore be based on different characteristics such as:

- Diabetic Patients
- Hypertensive Patients
- Patients with cardiovascular diseases
- Patients with pulmonary diseases

The main shortcoming of such grouping approach is the rigidity enforced by definitive characteristics of the groups. While expression of a medical condition as a binary or discrete variable does not prevent the expression of another condition for a patient, placing the patient in a single group disregards some of its features that could be similar to the patients in other groups. For instance, a diabetic patient with borderline hypertension could be placed in the “Diabetic Patients” group which actually ignores the fact that he/she also has some characteristics similar to the patients in “Hypertensive Patients” group. This loss of expressive power is considered as a drawback in applications that work with a large variety

of medical data. Since, for real world patient data, it is practically impossible for patients in a group to have absolutely no common characteristics with the patients from a different group, placing each patient into a single group would be unrealistic. This shortcoming could be alleviated at the expense of the highly specific association of each group with certain features in the dataset by removing the restriction that each sample must belong to only a single group. In machine learning, the clustering methods in which samples are allowed to be associated with multiple clusters are called fuzzy clustering methods [165]. By adapting a fuzzy clustering algorithm, it is possible to consider each patient's common features with those of patients belonging to different groups without losing the distinction introduced by the clustering process itself.

In this study, a variant of Fuzzy c-means (FCM) clustering, which is one of the most commonly used fuzzy clustering methods, is employed for the clustering task [165]. In FCM, the similarity measure used for clusters as well as the samples is the distance between them in the feature space. The association of each sample with clusters are described via membership grades, which can be considered as a measure of how well a sample fits in the characteristics represented by a given cluster.

The operation of FCM clustering algorithm is similar to that of k-means clustering with the main difference being that the samples now have c parameters that denote the membership grade for each of the c clusters, instead of a single parameter in k-means clustering. A brief, stepwise summary of FCM algorithm is provided on the next page with its concept and parameters explained in the following sections.

1-Set the number of clusters
2-Assign membership grades of each sample randomly
3-Calculate the coordinates of each cluster center (centroid)
4-Calculate the membership grades of each sample to minimize the objective function
5-Check if the convergence criterion is satisfied. If not, return to step 3.
6-Terminate the algorithm.

4.1.1. Number of clusters

The number of clusters is one of the most critical parameters in not only FCM but practically any clustering approach [165]. Generally, the number of clusters in FCM is defined by the user as the agglomerative methods such as dendrogram are not suitable due to the indefinite boundaries of each cluster and non-binary membership grades of the samples. Therefore, in most of the practical applications of FCM, the number of clusters are determined and possibly optimized according to the designer's choice, expertise and evaluation. In this study, the optimal number of clusters is determined by analyzing the performance of risk estimation module for cluster numbers c=1:10. The result of these analyses showed that the optimal number of clusters for this study is c=4.

4.1.2. Centroid

Centroid of a cluster corresponds to its center point, expressed in the form of feature based coordinates. In FCM, the boundaries of clusters are indeed fuzzy with no restriction imposed on the clusters' volume. Therefore, theoretically it is possible for all of the clusters to span entire sample space, albeit with diminishing density. It is, however, still possible to define a center point or centroid for a cluster which represents the weighted average of its members' coordinates in the feature space. The centroid of a cluster is calculated according to Eq 4.1.

$$c_k = \frac{\sum_{i=1}^{n} w_{ik}^{m} x_i}{\sum_{i=1}^{n} w_{ik}^{m}} \tag{4.1}$$

where c_k is the centroid of cluster k, n is the total number of samples, w_{ik} is the membership grade of sample x_i to cluster k and m is the fuzzifier, the parameter that determines on the fuzziness of the clustering process. For larger values of m, the clustering process will be fuzzier, resulting in a more homogeneous distribution of membership grades and vice versa.

4.1.3. Membership grade

In FCM clustering, the membership grade of a sample indicates how much does the sample belong to the given cluster by means of its similarity to other samples associated with the cluster. Due to the fact that samples may, and almost always do, belong to multiple clusters,

the membership grade of the samples are calculated with respect to the centroid of each cluster, as given in Eq. 4.2.

$$w_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_k\|}{\|\boldsymbol{x}_i - \boldsymbol{c}_j\|}\right)^{\frac{2}{m-1}}} \tag{4.2}$$

where $\|\boldsymbol{x}_i - \boldsymbol{c}_k\|$ denotes the Euclidean distance between sample x_i and centroid c_k in terms of feature coordinates with the same notation used for x_i and c_j.

In some fuzzy clustering algorithms, the only constraint on the membership grades was the limits on the values they could take individually, which is given in Eq. 4.3.

$$0 \leq w_{ik} \leq 1 \tag{4.3}$$

The FCM clustering used in this study utilizes a second constraint on membership grades. With the constraint given in Eq. 4.4, the membership grades are normalized, enabling an unbiased distribution among all samples. Considering how the clustering results will be handled in the risk estimation module, the choice of FCM as clustering method is considered to maximize the utilization of the information uncovered by the clustering process.

$$\sum_{k=1}^{c} w_{ik} = 1, \forall i \tag{4.4}$$

4.1.4. Fuzzifier

The fuzzifier, denoted by m is arguably the most important parameter in FCM clustering [166]. It determines the fuzziness of the clustering process by modifying not only centroids and membership grades but also the objective function. The fuzziness of the clustering is proportional to m with the distribution of samples among clusters becoming more uniform with increasing m. Smaller values of m modify the clustering process to be crisper, resulting in a higher variance of membership grades for each sample. On the extreme case of m=1, FCM clustering essentially becomes identical to k-means clustering as the membership grades converge to either 0 or 1, indicating a fully crisp clustering. While the choice on m largely depends on the application with no formal value or tendency specified for most

applications, conventionally it is set to be 2 by default. However, in this study m is determined by an exhaustive search algorithm scanning the interval [1 5] with a step size of 0,01 with the objective of maximizing risk estimation performance. The performance of the risk estimation module is observed to be maximized for m=1,56. Therefore this is chosen as the fuzzifier value used in the proposed model. The exhaustive search and its results in terms of classification accuracy is presented in Section 5.

4.1.5. Objective function

The FCM clustering algorithm aims to minimize the objective function given in Eq. 4.5 by iteratively calculating the membership grades for each sample and consequently the centroids of each cluster.

minimize $$\arg\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \|\mathbf{x}_i - \mathbf{c}_j\|^2 \quad (4.5)$$

Minimizing this objective function can be interpreted as minimizing the distance between each sample and centroid weighted by the membership grades. This implies that, as long as the constraints given in Eq. 4.3 and 4.4 are satisfied, the membership grade w_{ij} must be large for small distances between x_i and c_j and vice versa. Since a small distance between x_i and c_j means that x_i strongly belongs to cluster j, the membership grade w_{ij} essentially becomes a measure of this belonging.

4.1.6. Convergence criterion

Although the original model developed by Dunn in 1973 did not explicitly provide a formal definition of converge criterion [167], there are two main convergence criterions that could be employed in a FCM clustering algorithm [168].

Sensitivity threshold

A sensitivity threshold which corresponds to a lower limit of weight change between two iterations could be employed as a convergence criterion. It is up to the designer to specify whether the threshold is set for the maximum of changes in all weights in the model or for the average of weight changes for all samples. In both cases, a threshold ε is determined so

that if it the change in weights is less than ε at any point between two iterations, the clustering algorithm is said to converge and the process is terminated.

Objective function improvement

Alternatively, the convergence criterion for the clustering process could be set as the minimum improvement in objective function. For this case, a threshold ε is determined so that if the improvement, or the decrease, in the objective function's value is less than ε between two iterations, then the convergence criterion is said to be satisfied, terminating the process.
Objective function improvement is chosen to be the convergence criterion for this study with the threshold $\varepsilon = 10^{-5}$.

4.1.7. Inputs

The clustering process comprises the first stage of the proposed model. Therefore, its inputs are actually the inputs of the model itself. The three main inputs for the clustering module are discussed in this section.

Unclustered data

The unclustered data is the unlabeled form of the normal input data, obtained directly from the dataset. Due to the dynamic design of the model, the clustering task is carried out using both the training and test datasets. This is a necessary implementation because of the random nature of the clustering process, requiring the whole sample space to be available for observation in order to perform as intended. It is important to note that this approach does not violate the separation of training and test datasets as no training or evaluation is performed at the clustering module.

The usage of training and test datasets together in the clustering step is called joint clustering. This joint clustering enables the system to adapt itself according to the overall dataset, alleviating the effect of possible bias in the dataset. This property is implemented to maximize the performance of the proposed model regardless of the input data provided. In many cases the dataset formed from medical records of patients may contain biased data. If

the samples with biased data were not distributed among training and tests sets homogeneously, the performance of the system could be affected. For example, if the dataset is formed from multiple centers, the dataset obtained from one hospital may have some bias due to different patient profile or calibration of equipment. This may yield slightly different results in the data obtained from that hospital when compared to that from another one. Unless the training and test datasets are hand-picked to provide an even distribution from each center specifically, the distribution of data from the two centers might be in favour of one center in training set and the other center in the test set. If a conventional, static risk estimator is assigned to perform the risk estimation task for such datasets, it would be trained with a relatively biased data due to the nonuniform separation of samples from different centers. Unlike such an estimator which would have an increased error in prediction due to biased training, the proposed model would perform clustering with the new data just as with the original dataset used in this study, thus providing a new training dataset for the second module. In the new dataset, the samples with a nonrandom bias would likely be grouped together regardless of their stroke information which enables the adaptive usage of classifiers in the risk estimation module. By this way, the bias in the dataset would be reduced at the clustering stage even if the biased data is considerably concentrated in either of the test and training datasets.

Number of clusters

The number of clusters must be specified before the initiation of clustering algorithm. Although the optimal number of clusters is determined to be c=4 for this study, the proposed model is able to work with a range of different number of clusters; c=1:10, for the sake of enhancing the reproducibility of the results.

Fuzzifier

The fuzzifier is the parameter that determines the fuzziness of the clustering process. Therefore, it is treated as another input to the clustering module. While the proposed model does work with any acceptable value of m, it performs best with m=1,56. For this reason, this input is not asked from the user in the final model, but instead, its value is specified beforehand. It is still possible to alter m from the source code of the model.

4.1.8. Outputs

The FCM clustering algorithm used in the study provides the membership grades of the samples as the default output. For n samples and c clusters, this results in an output matrix of size n x c. The proposed model introduces another output which corresponds to the cluster number that each sample has the highest associated membership grade, denoted by “maxc”. This output, alongside the membership grade output is combined and added to the original dataset as additional features. These additional features can be considered as an implicit evaluation of each patient considering all of their available medical information represented by the original input features. The effect of membership grades and maxc on the classification performance is analyzed by tests carried out considering three possible combinations, alongside the base case without any clustering information added. The possible combinations are considering:

- Only the membership grades
- Only maxc
- Membership grades and maxc

It is shown in Section 5 that the performance of the risk estimation module is maximum for the third combination. Therefore, both membership grades and maxc were included as additional features in the final model. By this way the clustered dataset, which has c+1 additional input features to the original input dataset is formed. For the 124-feature input set and optimal number of clusters, this corresponded to clustered dataset having 129 input features.

The clustered dataset is partitioned and rearranged according to the maxc value of the samples via a simple sorting algorithm in order to prepare it to be transferred to the risk estimation module. This is a necessary step due to the design of the risk estimation module which involves calculation of weights for each classifier individually for each subset determined by the maxc of its samples. After this step, training and test datasets are determined randomly using 10-fold cross validation. For the 3000 patient dataset, this corresponded to the training set consisting of 2700 samples while the test set having 300 samples for each fold. The training and test sets are stored as a single dataset at this stage and the samples belonging to each set are identified according to their output features which

is discussed further in Section 4.2. The clustered dataset is then passed to the risk estimation module for the adaptive calculation of weights assigned to the classifiers that is carried out by employing only the training part of the dataset. Hence, the cascaded operation between two modules is provided by the transferring of the clustering module's output to the risk estimation module's input with only a rearrangement of the samples.

4.1.9. Memory and Computational Cost

The clustering process introduces an additional memory cost on the system as the input data matrix is expanded with the addition of membership grades and maxc features. This additional cost is, however, negligible as the original dataset already contains 124 input features. For the proposed model, the clustered dataset introduces 5 new features and therefore has 129 input features. This corresponds to 15000 additional numerical values in the resulting matrix containing 3000 samples. Considering that the numeric values are stored in double-precision format with 8 bytes allocated for each value, the total memory cost of additional data is calculated as 120 kB. This memory cost is negligible considering the capabilities of modern computers. Furthermore, the relative increase in memory cost due to clustering is only 5/124 or approximately 4,03%. Employing a crisp clustering technique or disregarding membership grades could result in a lower increase in memory cost by a factor of 5, however, the performance of the risk estimation module would be lower. Therefore, in the final model, the clustered data is formed by adding both the membership grades and maxc values to the input data.

Employing FCM clustering as a preliminary step for risk estimation naturally results in an increased computational cost when compared to performing risk estimation directly. The computational complexity of the FCM clustering algorithm used in this study is $O(nc^2p)$ where n is the total number of samples, c is the number of clusters and p is the total number of features [169]. Empirically, this corresponds to the clustering process to take no more than 2 seconds with n=3000, c=6 and p=124 when carried out on an Intel Core i7 2,2 GHz machine running Ms Windows 10 Home operating system. With a relatively powerful machine, it is possible to achieve near-instantaneous clustering even with a large number of samples or features. Therefore, the additional computational cost introduced by the clustering module is deemed minimal considering the improvement provided in prediction performance as presented in Section 5.

4.2. Risk Estimation Module

The risk estimation module comprises the second half of the proposed model with its input cascaded to the output of clustering module. In this module, the estimation of patients' stroke-related risks is carried out considering three stroke types, alongside the general stroke risk. The risk estimation is implemented by using model averaging methods to form an ensemble of classifiers for each of the 4 classification tasks, identified by the type of stroke.

Model averaging is the general strategy of reducing the prediction error by combining a number of different predictor models. The methods employing the model averaging strategy are known as ensemble methods in machine learning literature [170]. Ensemble methods aim to produce an optimal predictive model by combining several predictors or classifiers, referred to as base models in this context. Instead of seeking the best predictor model that is fit for a given task, which may not always exist, ensemble methods consider a number of models and combine their outputs to obtain a higher performance than these base models. Empirically, ensemble methods are shown to produce better results than any of their base models individually [155] even without complicated model averaging techniques in most cases. Therefore, their usage can be preferable in many tasks for which a definitively optimal model does not exist or is practically impossible to be created. Finding and optimizing such optimal models might be considerably tedious and may often result in dead ends unless the sample space and task itself are very well defined. Forming an ensemble of existing models that were proven to be somewhat successful and then optimizing the ensemble as a whole could be preferable for such cases.

The reason why ensemble methods tend to provide better performance is not directly related to the success of the base models but instead to their errors. To put more clearly, ensemble methods perform better than their base models because different base models will not usually make all the same errors on the samples. Unless the majority of base models' errors are somehow correlated, the errors made for most of the samples by some of the base models will be outweighed by the correct predictions made by the remaining base models. While being an empirical proof, this phenomenon forms the basis of the generally superior performance of ensemble methods. Due to their high performance, enhanced by their relative flexibility in optimization, ensemble methods are common top performers in artificial intelligence competitions [171].

The usage of multiple base models to perform a given task naturally results in increased computational load for ensemble methods. The computational load of an ensemble can be critical especially if a real-time operation is desired or the computational load of base models are already considerable, such as in image processing applications. However, for this study, the computational load is light enough to be of least concern considering the complexity of algorithms used in the proposed approach and offline operation. Therefore, the ensemble methods could be safely implemented for the risk estimation module.

The operation of this module is illustrated in Figure 4.2 with the steps explained in following sections. It should be noted that blocks representing dataset separation and the clustering module is also included in this figure in order to present a complete picture of the model operation. The operation of the proposed model up to stroke feature addition process was already explained in Section 4.1, therefore it will not be discussed again in this section.

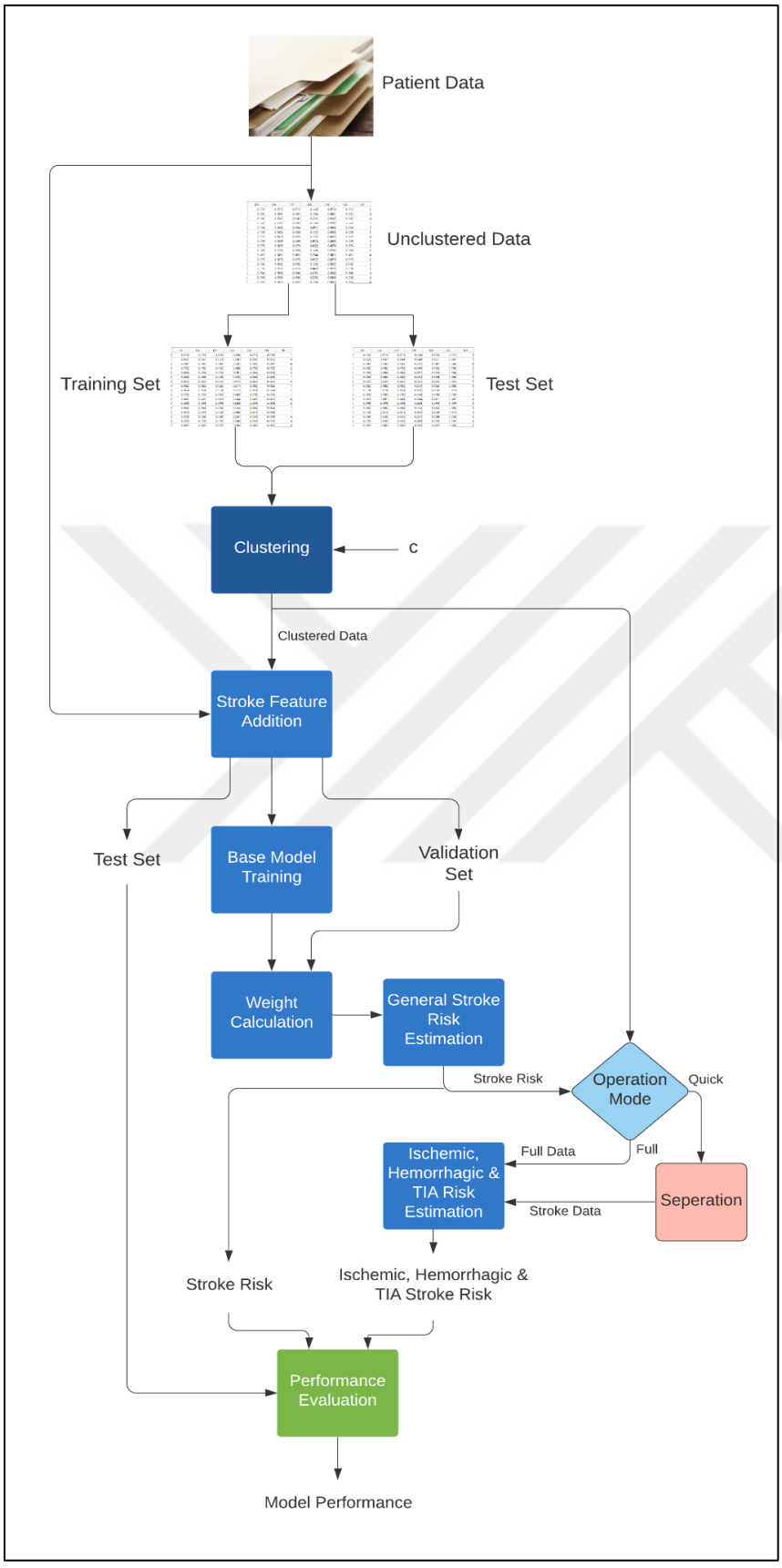


Figure 4.2. Flowchart summarizing the operation of risk estimation module

4.2.1. Training and validation

The base models comprising the risk estimation ensemble are classifier models that require supervised training to function properly. For this reason, the stroke information, expressed as 4 output features, was added to the clustered dataset at the input of the risk estimation module. For the training of the model, 10% of available data is spared for the test set. The output features for the test set are set to "-1" without separating its samples from the clustered dataset. By this way, the unitary structure of the input set is preserved, enabling the usage of structurally identical algorithms for training and testing of the model alongside making new predictions using it. The true stroke information of the test dataset is still stored separately to enable performance assessment after training. The training of the base models in the risk estimation module is carried out using the training set obtained as a result of this procedure.

The ensemble used in the study consists of 4 base models for each stroke type, resulting in 16 models to be trained. These models are selected from an initial list of 25 different classification algorithms, determined as the result of a comprehensive review of literature considering the most successful classifiers used in medical risk prediction tasks. Later, each of the classification algorithms are implemented and tested on 4 stroke types and for each stroke type, the most successful 4 methods were determined to be used in the study. The success criterion used in determining the methods was the accuracy-AUC product or AAP which is simply the linear product of validation accuracy and AUC for each classifier. The base models used for each stroke type in the study is provided in Table 4.1. Each of these models are discussed further in Section 4.2.3.

Table 4.1. Classifiers used as base models for stroke risk estimation module

	Hemorrhagic Stroke	Ischemic Stroke	TIA	General Stroke
1st Method	Standard Tree	Standard Tree	Simple Tree	Standard Tree
2nd Method	Cosine kNN	Fine kNN	Fine Gauss. SVM	Boosted Tree
3rd Method	Weighted kNN	Cosine kNN	Minkowski kNN	Bagged Tree
4th Method	Linear SVM	Boosted Tree	Balanced kNN	Balanced Gauss. SVM

It is observed from Table 4.1 that there are a variety of different classifier families such as SVM, kNN and decision trees are deemed suitable base models. Furthermore, for all 4 types of stroke, the highest performing classifiers generally belong to two of these families,

highlighting each stroke type has its unique predictor-response characteristics that is similar, but not identical to other stroke types.

The base models are trained using 10 fold cross validation on the training dataset which corresponds to each training subset containing 2430 samples. Since the features are needed to be provided to each base model in a specific order, the number of clusters that affect the clustering-related features are also considered as an input parameter for base model training. Therefore, for different values of c, different trained classifiers are obtained. The validation sets which are formed during 10 fold cross validation are then used to determine the accuracy of each base model for the given training set. The validation accuracies are calculated for each cluster separately, resulting in an accuracy matrix of size 4 x c. This is done because of the fact that the performance of each base model is unlikely to be equal for all clusters and some models perform better in some clusters. By this way, it is made possible to exploit this observation and pave the way to cluster-based optimization of the weights associated with each base model. As a result, a model having higher accuracy in predicting a given type of stroke risk for a cluster is given a larger weight while the same model may be given a smaller weight for another cluster since its accuracy for that cluster turned out to be lower.

The weights of the base models with respect to stroke types and clusters are determined to be inversely proportional to the ratio of incorrect predictions instead of being proportional to the accuracy which would result in a rather soft assignment. The mathematical expression of the weight assigned to base model i for cluster k is given in Eq. 4.6.

$$w_{ik} = 1/(1-a_{ik}) \tag{4.6}$$

where w_{ik} and a_{ik} are weight and validation accuracy associated with i^{th} base model and k^{th} cluster, respectively. In order to prevent improperly large weights, w_{ik} is forced to saturate at 10^6 if its calculated value exceeds this limit which corresponds to an accuracy of 99,9999%.

The proposed model allows the updating of the weights whenever new data with known output features is provided as input. In this case, unless manually run, the training of 16 base models is skipped and the new data is used as the new validation set that is combined with the original validation set as long as the initial training dataset is not erased from memory.

This allows the model to maintain its high performance even when the new dataset has considerably different characteristics in terms of sample distribution or bias. The high performance is maintained as long as at least one of the base models remains moderately successful in validation step thanks to the novel implementation of weights in the ensemble. Thus, by skipping the training stage, the main computational bottleneck, a considerable reduction in computational cost and training time is provided when working with new data which is expected to be frequent considering the main aim of the model. Therefore, the ensemble learns how much each base model should contribute to the decision making process without using computationally expensive secondary training algorithms such as stacking [172].

4.2.2. Testing and risk estimation using new data

The weights calculated after the validation step are stored in a 4 x c matrix which is obtained by application of Eq. 4.6 to the accuracy matrix. It should be noted that, for each stroke type, there is an associated weight matrix that contains the weights associated with the base models used to calculate the given stroke risk. The data of the patients for which the stroke risk is to be estimated, which were marked with "-1" on their output features, is then used to estimate the general stroke risk first, using the associated weight matrix. The general stroke risk of patient j is calculated according to Eq. 4.7.

$$G(j)=\frac{\sum_{i=1}^{4} Gb_i(j)*w_{Gik}}{\sum_{i=1}^{4} w_{Gik}} \quad (4.7)$$

where G(j) is the general stroke risk of patient j for which maxc(j)=k, $Gb_i(j)$ is the output of i^{th} base model for general stroke risk estimation and w_{Gik} is the weights associated with Gb_i and cluster k.

G(j) is a continuous variable within [0,1], representing the estimated general stroke risk of patient j which is regarded as one of the four main output parameters of the proposed model. However, in order to evaluate the model's performance objectively, the output parameters must also be available in the original format, which is binary for this study. Therefore, Gr(j) the binary representation of general stroke risk of patient j is obtained using Eq. 4.8.

$$Gr(j)=\begin{cases}1 & if\ G(j)\ \geq 0{,}5\\ 0 & else\end{cases} \tag{4.8}$$

This transformation is actually a x axis shifted version of binary step thresholding function and is illustrated in Figure 4.3.

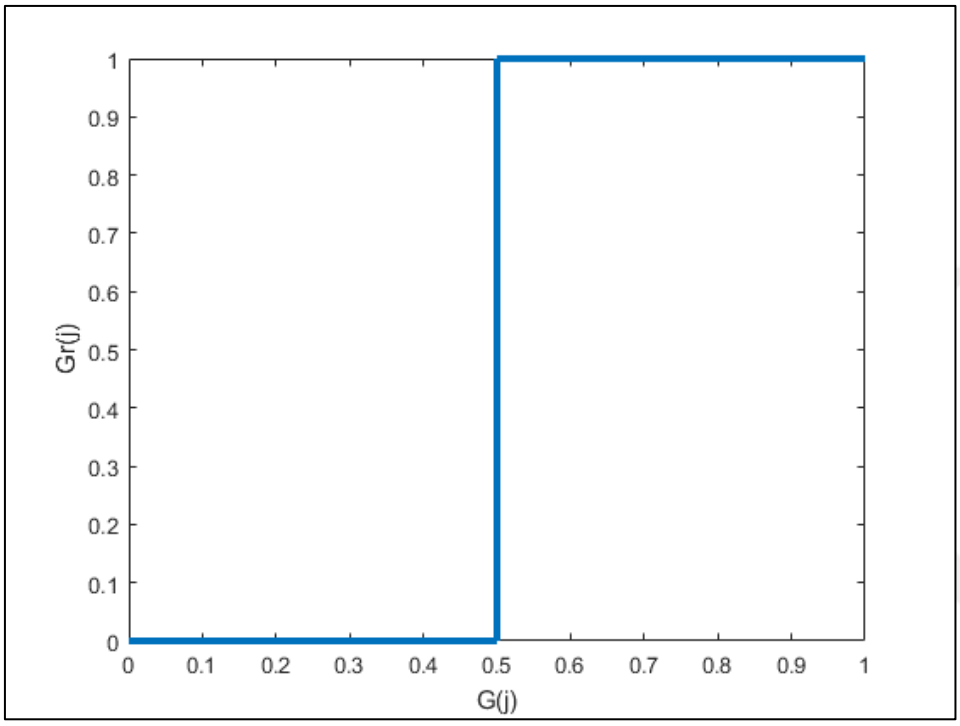


Figure 4.3. Shifted binary step thresholding function

Since the binary representation of the stroke risk has limited representative power, the continuous variable G(j) is included in the output of general stroke risk estimation function alongside Gr(j). For performance evaluation, which is mainly carried out based on the accuracy of the final model, Gr(j) is compared with the true values of the output feature "general stroke" for each sample in the test dataset.

The result of general stroke risk estimation is also used to determine the patients for which the risks of hemorrhagic stroke, ischemic stroke and TIA will be calculated depending on the operation mode selected by the user. If the operation mode is selected as "quick", the patients with G(j) = 0 will be separated from the dataset that is provided to the input of IHTRE block. In other words, the patients for which the risk of general stroke is estimated to be zero, are considered at little to no risk for all types of stroke in quick operation mode.

For these patients, the output features “hemorrhagic stroke”, “ischemic stroke” and “TIA” are automatically set to zero. The remaining patients’ data is then used to calculate the risk of hemorrhagic stroke, ischemic stroke and TIA for these patients. This results in a reduced workload and quicker completion of IHTRE process as a smaller subset of the original data is provided. The relative magnitude of these improvements is empirically observed to be proportional to the ratio of patients with G(j) = 0 to total number of patients in the dataset. The exact relationship between time and memory complexity of IHTRE process and sample size, however, depends on each of the base models used in the process and is outside the scope of this study. The disadvantage of this mode is the fact that for the patients that are separated from the dataset, the risk of hemorrhagic stroke, ischemic stroke and TIA are generally underestimated. This may also result in error amplification should there be false negatives with respect to general stroke risk in the dataset. For these false negative samples, one or more of the three types of stroke risk may be, and usually are nonzero yet they are automatically assumed to be zero by the model which results in further misestimation. In summary, this operation mode provides decreased operating time as well as memory and computational load for the proposed model at the expense of slightly lower accuracy. The quick mode is not used in model testing process as it does not utilize the full potential of the proposed model.

The other operation mode option is “full” mode and is the default operation mode. In this mode, the separator is skipped which results in the whole dataset is provided to the IHTRE block after general stroke risk estimation. Regardless of the operation mode, the IHTRE block estimates the risk of hemorrhagic stroke, ischemic stroke and TIA for the samples provided at its input. The risk estimation process for these 3 types of stroke is not different than those for the general stroke. In this block, 4 different base models are employed with their associated weight matrices that were generated in the validation process to estimate the risk for each stroke type. The ischemic stroke, hemorrhagic stroke and TIA risk for patient j are estimated using Equations 4.9 to 4.11, respectively.

$$I(j)=\frac{\sum_{i=1}^{4} Ib_i(j)*w_{Iik}}{\sum_{i=1}^{4} w_{Iik}} \tag{4.9}$$

$$H(j)=\frac{\sum_{i=1}^{4} Hb_i(j)*w_{Hik}}{\sum_{i=1}^{4} w_{Hik}} \tag{4.10}$$

$$T(j)=\frac{\sum_{i=1}^{4} Tb_i(j)*w_{Tik}}{\sum_{i=1}^{4} w_{Tik}} \quad (4.11)$$

Here I(j), H(j) and T(j) are the estimated ischemic stroke, hemorrhagic stroke and TIA risks of patient j for which maxc(j)=k, respectively. $Ib_i(j)$, $Hb_i(j)$ and $Tb_i(j)$ denote the output of i^{th} base model for ischemic stroke, hemorrhagic stroke and TIA risk estimation while w_{Iik}, w_{Hik} and w_{Tik} are the weights associated with Ib_i, Hb_i and Tb_i, respectively.

The estimations I(j), H(j) and T(j) are also used to obtain their binary counterparts Ir(j), Hr(j) and Tr(j) using the shifted binary step thresholding function given in Eq. 4.8 in a similar manner to Gr(j) to make accuracy evaluation possible. As a result of this process, the output of the IHTRE step has six variables, three of which represent the estimated risks of three stroke types while the remaining being their binary counterparts.

Joined by the general stroke risk feature predicted earlier, all four output features predicted by the model are available after the completion of IHTRE process. For the testing of the proposed model, the binary versions of these output features are used in the performance evaluation step which consisted of calculation of the accuracy for four stroke types. While other performance assessment criterions such as AUC or F-scores could also be selected for performance criterion, accuracy is believed to be a more universal measure of performance. Furthermore, equal distribution of test and control group patients in the dataset ensures to preserve the objectivity of this performance evaluation criterion which could otherwise yield a biased evaluation. Therefore, accuracy is selected as the main performance criterion while other measures could still be calculated with only slight adaptations in the performance evaluation block. The calculation of accuracy is carried out using the binary version of Rand index, with its expression for a general output feature Y is given in Eq. 4.12 [173].

$$R_Y = (TP_Y+TN_Y)/(TP_Y+TN_{Y+}FP_Y+FN_Y) \quad (4.12)$$

where R_Y is the Rand index accuracy with respect to feature Y; TP_Y, TN_Y, FP_Y and FN_Y are the number of true positives, true negatives, false positives and false negatives predicted by the model with respect to feature Y, respectively. The representation of possible features for

this study is G for general stroke, H for hemorrhagic stroke, I for ischemic stroke and T for TIA.

The performance evaluation is only carried out with the presence of test data or in the absence of new data in order to obtain the accuracy values presented in Section 5. The steps for using the proposed model to carry out risk estimation on new (unknown) data is identical to those described in this section with two main differences. Firstly, the user can select the operation mode as either "quick" or "full" contrary to fixed "full" mode during testing. Secondly, while the binary versions of risk estimations are still provided at the output of the model alongside original estimations, they are not used in performance evaluation which is redundant for normal operation. Furthermore, if deemed sufficient by the user, it is possible to perform only general stroke risk estimation which takes around a quarter of the time required for full operation, since the general stroke risk estimation step is separated from IHTRE step by design.

4.2.3. Base models used in risk estimation module

There are four different base models for each stroke type employed in the ensemble as demonstrated in Table 4.1. These base models are classification algorithms that could also be used as standalone estimators of a single output feature. Therefore, they exhibit the common properties of machine learning based classifiers. While there are 16 individual base models comprising the ensemble used in risk estimation module, it should be noted that some models such as Standard Tree are used in estimating multiple types of stroke. As a result, there are 12 unique base models used in the ensemble that belong to three different classifier families. In this section, these models which were trained individually in the training stage of risk estimation module will be presented with respect to the classifier families they belong.

Decision tree models

Decision tree models are nonparametric models that employ conditionally followed steps represented by the branches to predict the class or value of the response variable from the observations of the samples. The branches that symbolize the association of features with possible class labels are terminated by leaves of the decision tree. If the values that an output can take is limited, the decision tree is called a classification tree with the outputs generally

representing different classes. On the other hand, if the output is a continuous value that does not represent discrete classes, the model is called a regression tree. In that case the leaves do not represent a single value but rather an interval of values that a given output can take. The decision tree models used in the study are classification trees since the base models are used as classifiers in the ensemble.

The branches of a decision tree are generally furcated, with nodes among them. These nodes are associated with each of the input features and the path followed during the decision making process is drawn depending on the values of these features. The decision trees can be divided into two depending on the number of features considered at each node [129]. In univariate trees, only a single feature is considered in decision making at each node. At each node, a comparison of a given input feature is carried out with one or more thresholds and depending on the result of these comparisons, the branch to be followed is selected. For univariate trees, each node encountered during the decision making process, involves a different input feature, until a leaf is eventually reached. In a multivariate tree, at each node a unique combination of multiple input features is considered and the result of this combination, generally expressed as a function, is used in decision making. Due to the large number of input features considered, all tree models used in the study are determined to be multivariate trees.

The order of features to be encountered in the process is determined during the training stage of the tree which involves selection of a representative quality metric. A number of different metrics exist such as Gini impurity or information gain. For the tree models implemented in the study, Gini impurity index is selected as the splitting quality metric. Gini impurity index represents the probability of a randomly drawn sample from the dataset would be wrongly labeled if the classification was carried out randomly while preserving the distribution of each class among the dataset. Impurity, treated as a continuous variable within [0,1] refers to the degree of presence of minority samples among a sample set. Zero impurity for a set means that all the samples belong to a single class that is not necessarily the correct class. Hence, zero impurity may either correspond to entirely correct or entirely incorrect classification for a given set. An impurity equal to 1 means that for that set, there is no majority samples associated with any class label and the samples are randomly distributed across various classes. In the presence of c different classes, the Gini impurity associated with a given set of samples p is given in Eq. 4.13.

$$\text{Gini}= \sum_{i=1}^{c}(p_i \sum_{j\neq i} p_j) = \sum_{i=1}^{c}(p_i - {p_i}^2) = 1 - \sum_{i=1}^{c} {p_i}^2 \quad (4.13)$$

where p_i is the fraction of samples in p that belong to class i.

The trees are constructed by starting with a single node called the root node and performing partitioning recursively until a sufficient number of branches and leaves are obtained to represent the features and possible classes accurately. This learning method is considered as a greedy algorithm as it aims to obtain a well performing model by optimizing its stages instead of the model as a whole.

Decision trees are considered among structurally simpler machine learning algorithms. Due to their intelligible nature, they are relatively easy and to construct and train, generally requiring a small number of training data to achieve decent performance. They are, however, very sensitive to the changes in the training data and they occasionally tend to overfit which results in unnecessarily complex tree structures. For these reasons they are rarely used as standalone classifiers in modern applications. Ensembles of decision trees can also be constructed which is a popular approach to alleviate these shortcomings of standalone decision trees. Different ensemble methods can be applied to combine multiple decision trees which results in a model that is usually named after the ensemble method used. Examples of such models include boosted trees and bagged trees, both of which are employed as base models in the study. Therefore, the risk estimation module implemented within the proposed model can be considered as an ensemble of ensembles combined with various individual classifier structures as well.

There are two types of single-tree base models used in the ensemble that are characterized by the maximum number of splits allowed during their formation. These types are Simple Tree and Standard Tree. It is worth noting that for each stroke type, only one of them are present in the ensemble associated with the stroke type. It should further be noted that while there may be base models of same type for different types of stroke, these models are not shared. In other words, there are certain base models used for estimation of different types of stroke events that have similar or identical properties to each other. However, they are trained separately with different response variables, resulting in entirely different trained classifier structures.

Simple tree

The simple tree model constructed in the study has its maximum number of splits limited to 4 which results in a considerably compact model. As a multivariate tree, this resulted in a large number of features to be considered at each node. This is balanced by imposing no limits on the branching factor; that is the number of splittings or new branches performed at each non-leaf node. Simple trees, thanks to their low number of branches are more resistant to overfitting however, they are weak learners, that is they have a large bias and tend to underfit. These disadvantages of simple tree models are neutralized by employing more complex models within the ensemble while keeping a fair accuracy.

Standard tree

The standard tree model is formed using the same methodology described earlier. For this model, the maximum number of splits is limited to 20. This resulted in a stronger learning of the training set when compared to the simple tree structure. By keeping the maximum number of splits relatively low, overfit is largely avoided while the bias is reduced. Thanks to these balanced characteristics, standard tree models are considered as flexible models that can achieve decent accuracy in different classification tasks. For this reason, standard tree is the most frequently used base model type in the ensemble, comprising 3 of the 16 individual classifiers.

Ensembles of trees, also referred to as forests could also be employed as base models in an ensemble in addition to their usage as separate classifiers. Bagged trees, boosted trees and rotation forest classifiers are among examples of ensembles of trees [174]. Among these, bagged and boosted tree models are used in the ensemble.

Bagged Tree

Bagged tree is the ensemble of trees formed by bagging or bootstrap aggregating. Bagging is a voting based ensemble method in which the base models are trained using overlapping subsets of the training dataset. In this case, each subset used in training of a base model has some unique samples alongside samples that are also present in other subsets. These subsets are formed by the random sampling of the original dataset with replacement. It should be

noted that unless additional conditions are enforced, this random sampling may result in different frequency of each sample with some samples not being present in any subset at all.

The training of each base learner in the ensemble is carried out separately using the associated subset. For bagged trees, this corresponds to growing each tree individually using its subset, as described earlier. After the ensemble is formed, the output of each base model is brought together at the output stage, which is used to determine the output of the ensemble either by averaging or voting. Averaging is generally used in regression models where the output is a continuous variable. For binary classification tasks, which is the case for base models used in the study, voting is the typical method in determining the ultimate output. In this case, the class label that is predicted by highest number of the base models is selected as the output of the ensemble.

The bagged tree model used in the study consisted of base trees that are similar to the aforementioned standard tree model in terms of complexity. The number of maximum splits is limited to 20 for each tree. The bagging is performed by dividing the training dataset of 2700 samples into subsets containing 1500 samples each, corresponding to each subset having roughly 43% of its samples as unique samples, calculated according to Eq. 4.14 [175].

$$E_m^n = (1 - e^{-n/m}) \tag{4.14}$$

where E_m^n is the expected ratio of unique draws when drawing n samples from a set of m units with replacement. The final output of the bagged ensemble is determined by voting.

Boosted Tree

In bagging, the subsets used to train each base model is generated randomly even with necessary precautions to ensure homogeneous frequency for all samples. This results in a rather random performance of bagged structures which is their main disadvantage.

Boosting aims to alleviate this disadvantage by implementing a different strategy in subset formation. In this case, the initial division of the original set is still performed randomly. Using one of these randomly generated subsets, the first base model in the ensemble is trained. Then, another subset is fed to the trained base model and its incorrectly classified

elements are retrieved alongside an equal number of correctly classified elements selected randomly. By this way, the second training set that will be used in training of the second base model is formed. With the second base model trained, the procedure is repeated on both base models to form the training set for the third model. With this recursive training and validation approach, all of the base models are iteratively trained with each base model utilizing the subsets formed depending on the operation of previous base models. Training the next base model based on the mistakes of previous learners was shown to result in reduced error rate when compared to bagged models of similar complexity [176].

The main difficulty in realizing a properly trained boosted model is the fact that unless the dataset is considerably large or the first few classifiers are at least moderately inaccurate, the training subset used in next classifiers diminishes rapidly. The application of boosting to a group of decision trees resulted in boosted tree structure which involves generation of trees with a limited number of branch nodes sequentially. Each tree is formed using the subset that employs a portion of validation sets from the previous trees. This formation of individual trees is no different than that of standard decision trees with the main difference between the base trees in the model is the usage of different training subsets.

The boosting method used for the boosted tree model is a variant of Adaboost, short for adaptive boosting as proposed by Freund and Schapire [177] that alleviates the diminishing subset problem by reusing the training set. In order to prevent overfitting for this case, the tree models are needed to be relatively simple. For this reason, the maximum number of splits allowed in base trees is limited to 12. Since the main training set is constant, generation of unique subsets to train each tree is now based on a probabilistic approach. At the beginning of the algorithm, the probability $p_1(x_i)$ of each sample x_i is equal to 1/n where n is the total number of samples in the dataset. The subset used to form the first tree is generated randomly, as in bagged trees. The training error ε_1 of the first model is then used to modify the probability of occurrence of each misclassified sample as described in Eq. 4.15.

$$p_2(x_i) = \frac{1-\varepsilon_1}{\varepsilon_1} p_1(x_i) \tag{4.15}$$

where $p_2(x_i)$ and $p_1(x_i)$ are the posterior and initial probabilities of occurrence of sample x_i, respectively. The probability of occurrence of correctly classified samples are held constant at each iteration. Therefore, the probability of occurrence of sample x_i on the subset used to form the j^{th} tree can be expressed as in Eq. 4.16.

$$p_j(x_i) = \begin{cases} \frac{1-\varepsilon_{j-1}}{\varepsilon_{j-1}} \, p_{j-1}(x_i) & if\ x_i\ is\ misclassified \\ p_{j-1}(x_i) & else \end{cases} \tag{4.16}$$

It should be noted that in between calculation of $p_j(x_i)$ and formation of j^{th} tree, the probability distribution p_j is normalized to satisfy $\sum_i p_j(x_i) = 1$. As a result of the alteration of $p_j(x_i)$, the chance of misclassified samples to appear in the next training subset is increased while that of correctly classified samples is decreased gradually due to normalization. The output of the boosted tree structure is determined by voting, similar to the bagged tree model. The model used in the study consists of 30 decision trees that are grown sequentially.

K-nearest neighbors models

k-nearest neighbors (kNN) is a family of nonparametric models that consider the k number of neighbors of each sample upon determining the output associated with that sample. kNN methods could be used in both regression and classification problems. In regression, the value of a given sample is calculated as a mean or median of its k neighbors while in classification, the class label is determined according to the labels of the sample's neighbors, mostly by voting.

The training stage of kNN algorithms does not actually involve a conventional model training that employs learning methods discussed in Section 2.2.3. Instead, at this stage, the distance between each sample in the dataset is calculated according to the distance metric that is determined by the type of kNN algorithm used. Then, for each sample whose class is to be determined, the k nearest samples with known class labels are used to assign a class label to that sample. In the classical version of kNN, the class label of a sample is determined by voting of its neighbors. The class with highest number of member neighbors is assigned to the sample. In order for the algorithm to function properly, k must be much smaller than the total number of samples. On the extreme case of k=1, which is specifically named

1-nearest neighbor classifier or fine kNN, all samples are assigned to the same class of their nearest neighbor.

kNN algorithms constitute the second most frequently used base model family in the ensemble with five different types employed as six different base models. Only the general stroke risk estimation did not feature any kNN algorithm.

Cosine kNN

Named after the distance metric it employs, Cosine kNN is considered among the most robust kNN classification algorithms [178]. Unlike the Euclidean distance, the cosine distance does not actually represent a measure of spacial distance but instead it represents the cosine of the angle between the vectors of two samples. Each sample in the dataset can be expressed as a vector pointing toward the location of that sample from origin in the sample space whose dimensions are defined as the features of the dataset. The cosine of the angle between two vectors can be used as a measure of similarity between them, which is referred to as cosine similarity [179]. Since the general approach in kNN is to define a measure of distance, instead of similarity, the concept of cosine distance is introduced in relation to cosine similarity metric. The cosine distance between two sample vectors **A** and **B** can be calculated as given in Eq. 4.17.

$$\text{Cd}=1-\text{Cs}=1-\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} \tag{4.17}$$

where Cd and Cs are the cosine distance and similarity values, respectively. Here, $\boldsymbol{A}\cdot\boldsymbol{B}$ is the scalar product of vectors **A** and **B** while $\|\boldsymbol{A}\|$ represents the magnitude of vector **A** over all dimensions.

The operation of cosine kNN algorithm used in the study is identical to classic kNN with k=10, the only difference being the distance metric. Neighbors of each sample are determined by exhaustive search which constituted most of the computational load of the algorithm. The class label for each sample is determined by the plurality vote of its neighbors.

Fine kNN

Fine kNN is a variant of classical kNN algorithms that is characterized by using only the closest neighbor in determining the class label of a given sample. Therefore, k is set to be 1 for this model. The fine kNN variant used in the study employs Euclidean distance over n=129 dimensions as the distance metric, which is given in Eq. 4.18.

$$||\mathbf{A}\text{-}\mathbf{B}|| = \mathrm{Ed}(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (4.18)$$

where ||**A-B**|| and Ed(**A,B**) are the representations for Euclidean distance between sample vectors **A** and **B** with their components in the i^{th} dimension as A_i and B_i, respectively.

Balanced kNN

One of the classical variants of kNN algorithms, Balanced kNN makes relatively few distinctions when compared to Fine kNN. The model used in the study considers k=10 neighbors for each sample, determined by the same distance metric as in Fine kNN, Euclidean distance.

Minkowski kNN

Minkowski kNN is named after the distance metric it employs, namely the Minkowski distance. The Minkowski distance could be considered as a generalization of Euclidean distance and Manhattan distance as it is possible to obtain these distance definitions with specified orders of Minkowski distance. The general expression of Minkowski distance of order p between n dimensional vectors **A** and **B** is given in Eq. 4.19.

$$\mathrm{Md} = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{1/p} \quad (4.19)$$

where Md is the Minkowski distance, A_i *and* B_i are the components of vectors **A** and **B** in the i^{th} dimension or feature, respectively. It should be noted that for p=1 the expression in Eq. 4.19 simplifies to Manhattan distance and p=2 corresponds to the well known case of Euclidean distance.

The Minkowski kNN base model used in the study employs the Minkowski distance of order p=3 as its distance metric. Furthermore, the number of neighbors to be considered is k=10, same as in Balanced kNN model.

Weighted kNN

In the kNN models discussed earlier, the weight of each neighbor's class in determining the class label of a given sample was equal. The Weighted kNN model introduces weights on each of the neighbors' vote in determining the class of the samples. The weights assigned to the neighbors are inversely proportional to the distance between the neighbor and sample while a variety of different expressions for weights exist such as inverse or squared inverse. With the usage of weights, the neighbors that are closer to the sample are given a higher contribution in the decision making process and the model is made more resistant to possible bias in the traning dataset [180].

The Weighted kNN model used in the study utilizes the same number of neighbors as most of the other models; that is k=10. Moreover, the distance metric employed for this model is Euclidean distance making Weighted kNN essentially a weighted version of the Balanced kNN model used in the study. Considering the large number of samples in the training dataset with binary classes for each task with a relatively low value of k, the neighbors of a random sample are expected to be located within a relatively small radius. Therefore, in assigning the weights, a slightly steep function is determined to be used in order to utilize the weights in decision making process more effectively. The function used to determine the weights associated with a sample x_i and its neighbors is given in Eq. 4.20.

$$w_{j,i} = \frac{1}{N_i\left(Ed\left(n_{j,i}, x_i\right)\right)^2} \tag{4.20}$$

where $w_{j,i}$ is the weight assigned to $n_{j,i}$; the j^{th} neighbor of x_i, $Ed\left(n_{j,i}, x_i\right)$ is the Euclidean distance between x_i and $n_{j,i}$ and N_i is the normalization coefficient which is calculated according to Eq. 4.21.

$$N_i = \sum_{j=1}^{k} w_{j,i} \tag{4.21}$$

The class label of x_i, y_i is determined by the weighted voting of $n_{j,i}$ as given on Eq. 4.22.

$$y_i = \sum_{j=1}^{k} w_{j,i} \cdot y_{j,i} \quad (4.22)$$

where $y_{j,i}$ is the label of j^{th} neighbor of x_i.

Support vector machine models

Support vector machines (SVM) are another family of learning models that are commonly employed in classification and regression tasks due to their robustness and relative simplicity [181]. They are also computationally less expensive than many robust models as the computational complexity of a standard SVM is $O(n^2k)$ where k is the number of support vectors that will be introduced later in this section [182]. While SVMs can also be employed in unsupervised learning to perform tasks such as clustering, the usage of SVMs in this study is limited to supervised classification hence they will be regarded as supervised classifiers hereafter. SVMs aim to perform classification by separating the samples from different classes by forming a hyperplane in a multidimensional coordinate system or space that the samples are mapped to. This space is either the normal or transformed version of the multidimensional feature space with dimensions corresponding to the features in the dataset depending on the type of the SVM.

If the samples are linearly separable, there is generally an infinite number of hyperplanes that can separate them. The hyperplane with most distance from the closest sample or samples is defined as the maximum margin hyperplane. The sample or samples that are closest to the maximum margin hyperplane are called support vectors, giving the method its name.

The classical version of SVM seeks a hyperplane that separates the samples correctly in the original feature space, a two dimensional case of which is illustrated in Figure 4.4. It should be noted that since a hyperplane is always of 1 less dimension than the coordinate system, the hyperplane in Figure 4.4 is actually a 1 dimensional line.

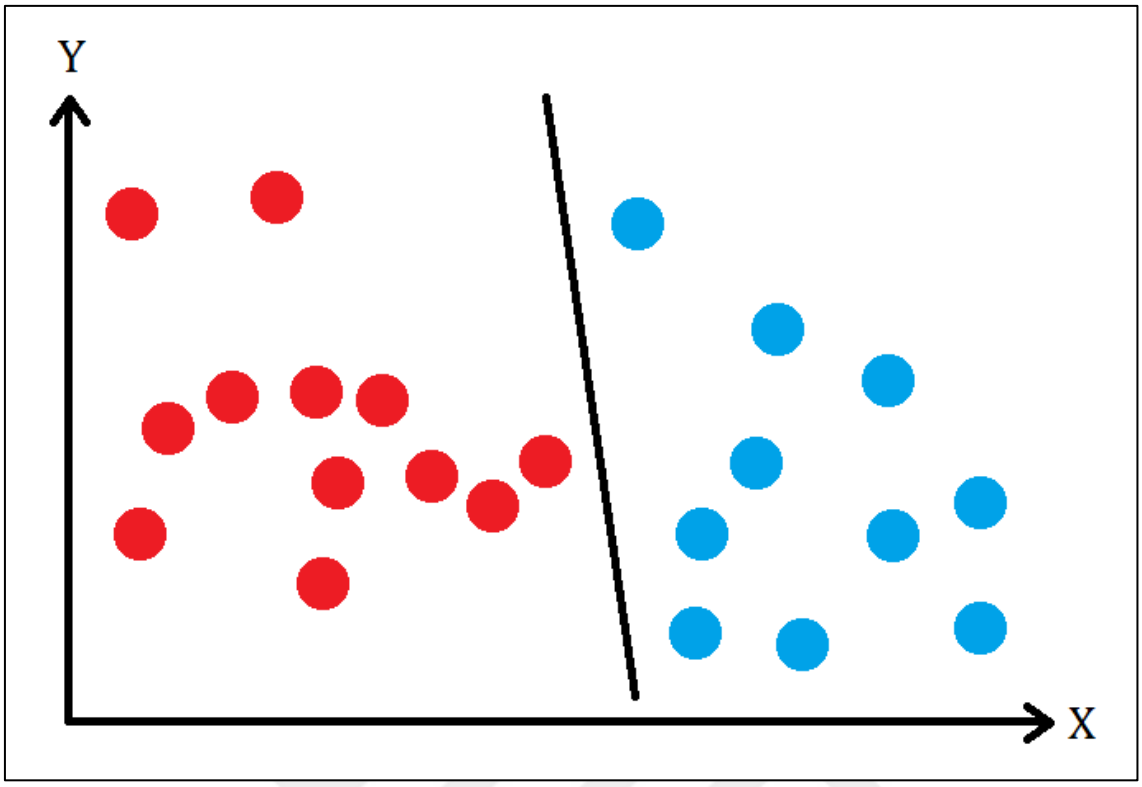


Figure 4.4. Separation of samples from two different classes by a SVM hyperplane (line)

The objective function of a m dimensional SVM used in binary classification of n samples is given in Eq. 4.21 [182].

$$\text{minimize}_{a_0,\ldots,a_m} : \sum_{j=1}^{n} MAX\{0,\ 1-(\sum_{i=1}^{m} a_i x_{ij} + a_0)\, y_j\} + \lambda \sum_{i=1}^{m} (a_i)^2 \qquad (4.21)$$

where $a_0,\ldots,a_m$ are the coefficients defining the m-dimensional hyperplanes $a_o x_1+a_1 x_2+\ldots+a_m x_m = M$ and $a_o+a_1 x_1+\ldots+a_m x_m = -M$ with M being the scaling factor which is chosen to be 1 for present case. x_{ij} is the component of j^{th} sample x_j on i^{th} dimension, corresponding to i^{th} feature and y_j is the class label of x_j. λ is the margin optimization parameter that is chosen to be unity for the models constructed in this study for a balanced cost function that emphasizes accuracy and margin equally.

In Eq 4.21, it should be noted that y_j is either +1 or -1, enabling the usage of the first term as a measure of error. With a proper separation, the samples resided on one side of the hyperplane would satisfy $a_o+a_1 x_1+\ldots+a_m x_m \geq 1$ while the samples on the other side of the hyperplane satisfying $a_o+a_1 x_1+\ldots+a_m x_m \leq 1$. Assigning +1 to the first side and -1 to second,

it is possible to fuse these two conditions to a general expression, satisfied by all correctly classified samples as given in Eq 4.22.

$(a_o+a_1x_1+\ldots+a_mx_m)\, y_i = ((\sum_{i=1}^{m} a_i x_{ij}) + a_0) y_j \geq 1$ (4.22)

Therefore, the first term in the objective function becomes zero whenever all the samples are classified correctly, resulting in a well trained SVM. Similarly, the second term in Eq. 4.21 is minimized by determining the smallest values for coefficients $a_0,\ldots,a_m$ while sustaining a high accuracy.

The training stage of the SVM consists of determination of coefficients $a_0,\ldots,a_m$ using the samples from the training dataset. The planes defined by the coefficients are later used in classification of new data. The ensemble in the risk estimation module contains three different SVM models; one linear SVM and two nonlinear versions of SVM that utilize kernel functions that will be explained alongside these base models.

Linear SVM

Linear SVM is the original and most elementary form of SVMs that are proposed as a remedy to linear classification problems. While their initial usage was limited to linear classification, the simplicity of Linear SVMs made them an attractive alternative in nonlinear classification problems as well. Theoretically it is not possible to achieve 100% accuracy by using a Linear SVM on a dataset that is not linearly separable. However, by sacrificing some accuracy, it is still possible obtain a stable and robust classification even with a high dimensional dataset such as the one used in this study [183].

The Linear SVM forms the dividing hyperplane essentially on the feature space which is m=129 dimensional for the base model used in the study. The training of the SVM is carried out considering the objective function given in Eq. 4.21 with j=1:2700. Since the dataset is not linearly separable, the algorithm is trained to maximize the soft margin, allowing a number of misclassifications to maximize the separation width. A box constraint that is inversely proportional to the number of support vectors and controls the maximum penalty applied to samples remaining inside the margin is introduced to prevent overfitting.

Fine Gaussian SVM

The samples in datasets used in practical applications is rarely linearly separable, limiting the usage of linear classifiers for most cases. While the separating hyperplane in a SVM is linear by definition, the linearity is not necessarily in the original feature space. It is suggested that any nonlinearly separable set of samples can mapped to an arbitrarily large dimensional space in which they become linearly separable [184]. Nonlinear SVMs exploit this idea by employing transformation functions called kernels to map the data into a transformed space, which is generally of higher dimension than the original space [185]. With feature transformation, a coordinate system in which the samples become linearly separable is expected to be found. Therefore, the samples are separated by a hyperplane formed within the transformed feature space without violating the general hyperplane definition. In other words, for the samples x_j that are not linearly separable, the nonlinear SVM seeks and utilizes a kernel transformation k() for which the transformed samples $k(x_j)$ are linearly separable. Assuming that the transformed space is generally of higher dimension than the original feature space, the general form of kernel function can be expressed as in Eq. 4.23.

$$k(\mathbf{x}) = (\mathbf{x},\mathbf{x}') \tag{4.23}$$

where **x** is the coordinate vector corresponding to original feature space and **x**' represents the additional dimensions introduced by kernel transformation. While there is no formally proven theory regarding the best kernel function tailored for a given problem, the general approach is to try and modify various kernel functions that were proven robust in literature. Polynomial, sigmoid and Gaussian kernels are among the most commonly used kernel functions in SVM applications [186]. Once a suitable kernel function is determined, the training data is transformed using this function and the determination of optimal separating hyperplane is carried out in a very similar manner to Linear SVM. The hyperplane for this case is determined in the transformed space using the objective function given in Eq. 4.21 only with a slight modification of replacing x_{ij} and y_j by their transformed counterparts $k(x_{ij})$ and (y_j). The expression for the hyperplane in the original feature space could be obtained by inverse transformation $k^{-1}(\mathbf{x},\mathbf{x}')$. However, this is neither necessary nor relevant for SVM to be trained or used in classification properly.

One of the main advantages nonlinear SVMs is that the improvement brought by kernel transformation comes at little computational cost. The mapping process by kernel transformation has a time complexity of O(n) which is dissolved within the complexity of the original algorithm. Furthermore, once the SVM is trained, classification of new data can be carried out with transformation of only the support vectors that represent the boundary data, resulting in faster prediction as well as a low memory cost [185].

The Fine Gaussian SVM model built in the study employs a Gaussian kernel function that does not explicitly transform the whole data in order to reduce computational complexity. The method forms a Gram matrix that contains the inner product of the kernel transformed versions of the sample pairs. It should be noted that Each element of the Gram matrix is calculated without explicitly performing the kernel transformation on each sample, which is known as the kernel trick [187]. Consequently, the elements of the Gram matrix can be obtained using Eq. 4.24.

$$G_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2,8}} \tag{4.24}$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are the vectors corresponding to i^{th} and j^{th} samples, expressed in feature space, respectively. In this model, a relatively low value of kernel scaling parameter (2,8) is chosen to provide a finer distinction of classes, giving the base model its name. Theoretically, the dimensionality of the transformed space is infinite for Gaussian kernel transformations. However, in practice, the number of dimensions is equal to the number of training samples which is 2430 with the usage of 10-fold cross validation on the 2700 sample dataset.

Balanced Gaussian SVM

The Balanced Gaussian SVM model is very similar to the Fine Gaussian SVM model described earlier. It uses a Gaussian kernel function to map the samples into a transformed space where they become linearly separable. Kernel trick was also employed to circumvent the computational complexity of explicitly transforming all samples. As in Fine Gaussian SVM model, the transformed feature space is 2430 dimensional since the base models are trained with the same dataset. The resulting Gram matrix is described in Eq. 4.25.

$$G_{ij} = e^{-\frac{\|x_i - x_j\|^2}{11}} \quad (4.25)$$

As can be seen in Eq. 4.25, a larger value of kernel scaling parameter is used when compared to Fine Gaussian SVM model. This results in a relatively weaker predictor that is more resistant to overfitting. Since much coarser distinction of classes is possible with higher kernel scaling factors, this base model is named Balanced Gaussian SVM emphasizing its balanced characteristics.

4.3. Graphical User Interface

After both modules are implemented and cascaded, the model is complemented with a graphical user interface (GUI) to provide a user-friendly operation. The GUI is designed to be easily understandable and usable keeping in mind the end user who may be a medical practitioner without in-depth knowledge of machine learning. Therefore, the GUI is devoid of any complicated terminology regarding machine learning or artificial intelligence while preserving the full functionality of the developed model. The design of GUI is carried out on Matlab's graphical design environment (GDE) named Guide.

Guide provides the designer the necessary tools to place a variety of interface elements such as pushbuttons, drop-down lists called popup menus and tables. A screenshot of Guide with the design in its later stages is given in Figure 4.5.

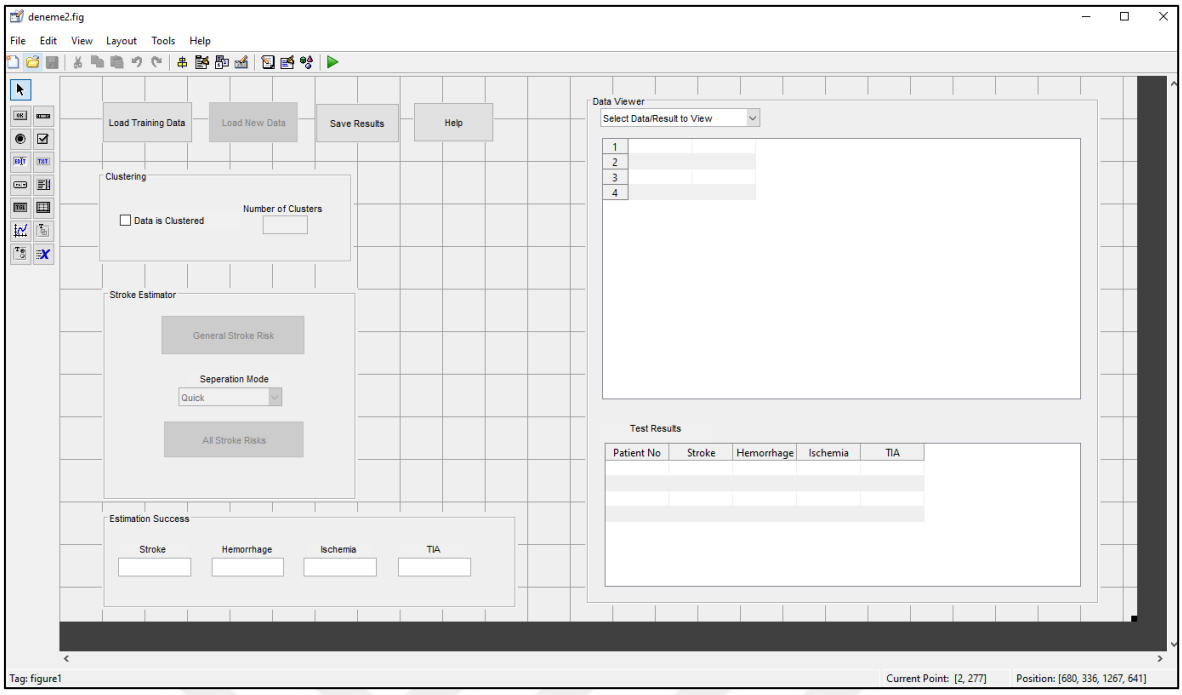


Figure 4.5. Guide interface with the design approaching its final form

The final version of the GUI consists of a number of pushbuttons, editable and non-editable textboxes, drop-down menus, table viewers and a non-editable checkbox, grouped depending on their function. It can be analyzed in four parts depending on their purposes as illustrated in Figure 4.6.

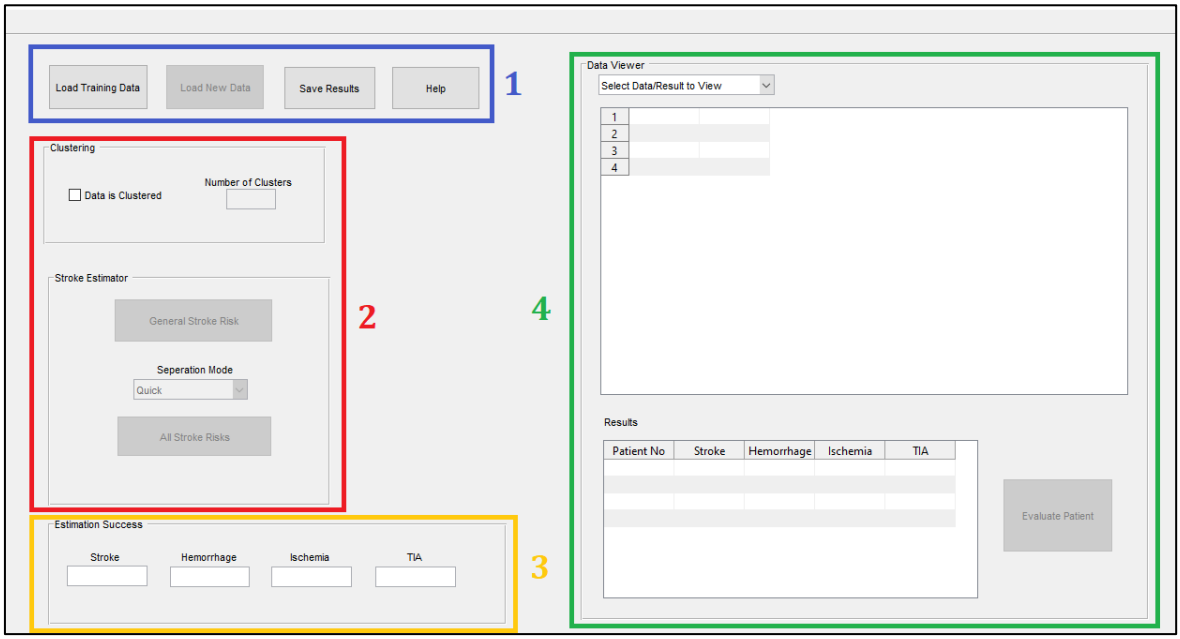


Figure 4.6. Final version of the GUI for the proposed model

4.3.1. Home button controls

The home button controls consist of four pushbuttons that are required for basic operation and understanding of the GUI. They are also the first group that the user will interact when using the GUI. The close-up image focusing on the home button controls is given in Figure 4.7.

Figure 4.7. Home button control group

Load training data

This button is used to load the data that will be used in training of the base models in the risk estimation module of the proposed model. The training data must be in a table format with at least 128 columns with first 124 corresponding to input features and the remaining 4 corresponding to the output features for the system to function properly. If the selected dataset contains exactly 128 columns, the model considers it as unclustered and enables selection of number of clusters as can be seen within the 2nd group in Figure 4.6. If the training dataset contains more than 128 columns, the model assumes that the dataset is already clustered and regards the additional columns as the clustering data. If a new training data is loaded while there is already a training data in memory, the existing data is replaced with the new training data and any result that is not saved is erased from memory.

Load new data

The second button in home button controls is the “Load New Data” button that is used for loading the new (or unknown) data without known output features. This new data is added to the end of the existing dataset in the memory. It is possible to add multiple new datasets one after another using this button multiple times. The number of columns in the new data must be greater than or equal to 124. If a dataset containing more than 124 columns are loaded, all elements of the excess columns are replaced with “-1” and the model assumes

that a test dataset is loaded. However, if the number of columns in the new data is greater than 128, the program throws an error, stating that the dimensions of the training and test data must agree.

This pushbutton is disabled (greyed out, unclickable) when the program is initialized. Upon loading a training data successfully, it becomes enabled, allowing the user to add as many new data as desired. When the risk estimation process is initiated using the 2nd group in Figure 4.6, "Load New Data" button becomes disabled again as it is not possible to modify the dataset with risk estimation process already running. Furthermore, this pushbutton remains disabled if the loaded training dataset is already clustered as it is not possible to perform clustering on training and new data separately.

Save results

This pushbutton is used to save the variables formed during the program operation in a file specified by the user. Since the workspace of GUI is separate from the base workspace of Matlab and is not directly accessible, the user needs to save the results using this function. Depending on at which stage of program operation it is pressed, this pushbutton saves the information given in Table 4.2. It should be noted that, if no training data is loaded at the time of save operation, the save file will be blank.

Table 4.2. The information saved by "Save Results" pushbutton

Saved Information	Operation Stage
Merged dataset containing training and test data, if available	After loading a training dataset Before general risk estimation
General stroke risk estimates	After general stroke risk estimation
General stroke validation accuracy	After general stroke risk estimation
Ischemic stroke, hemorrhagic stroke and TIA risk estimates	After all stroke risk estimation
Ischemic stroke, hemorrhagic stroke and TIA validation accuracy	After all stroke risk estimation

Help

The last button in home button controls is the “Help” button that displays a help document in the form of a large dialog box when pressed. This help document is provided in Figure 4.8.

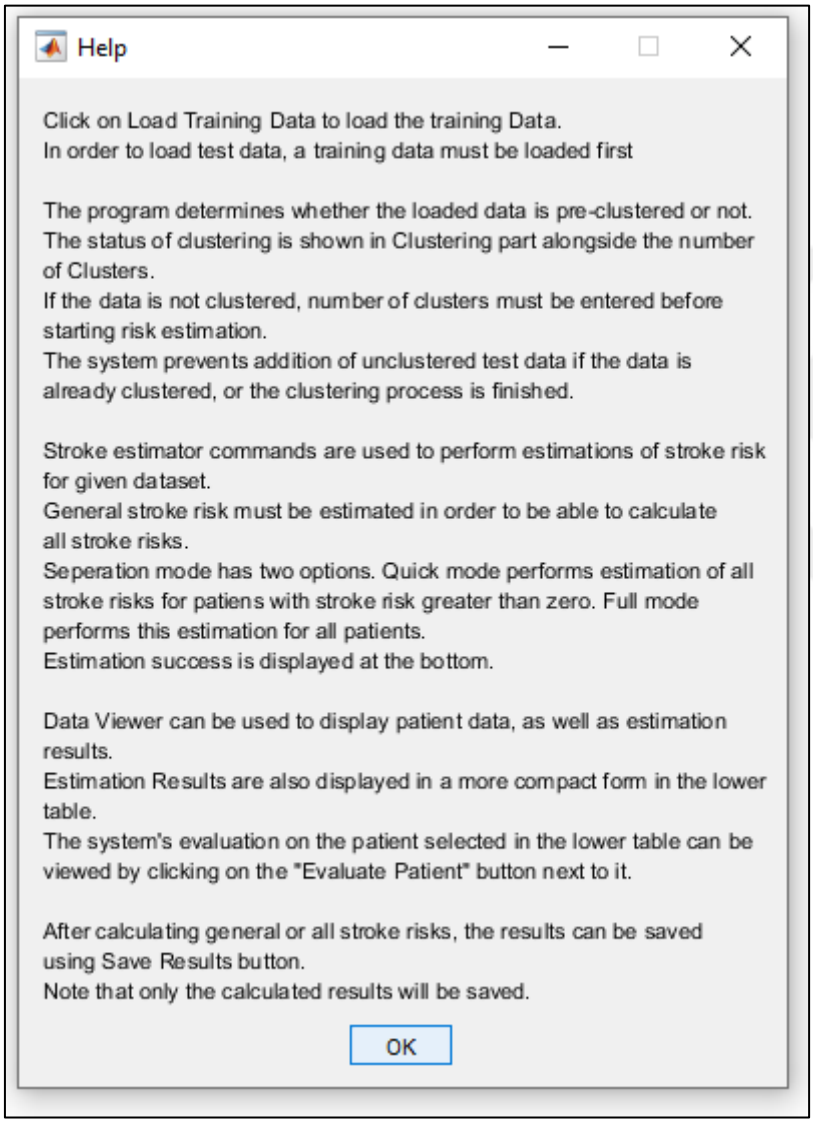


Figure 4.8. The help document displayed when the pushbutton “Help” is pressed

4.3.2. Clustering and stroke estimation controls

The main operation of the model is controlled by the components in the clustering and stroke risk estimation part of the GUI. Clustering and stroke estimation is carried out after the necessary datasets described in Section 4.3.1 are loaded. A close-up view of the Clustering

and stroke risk estimation controls are given in Figure 4.9. As can be seen from Figure 4.9, the clustering and risk estimation modules are graphically separated in parallel with their implementation as two separate modules in the proposed model. It should be noted that all components of this group start as disabled in program initialization and remain so until a valid training dataset is loaded.

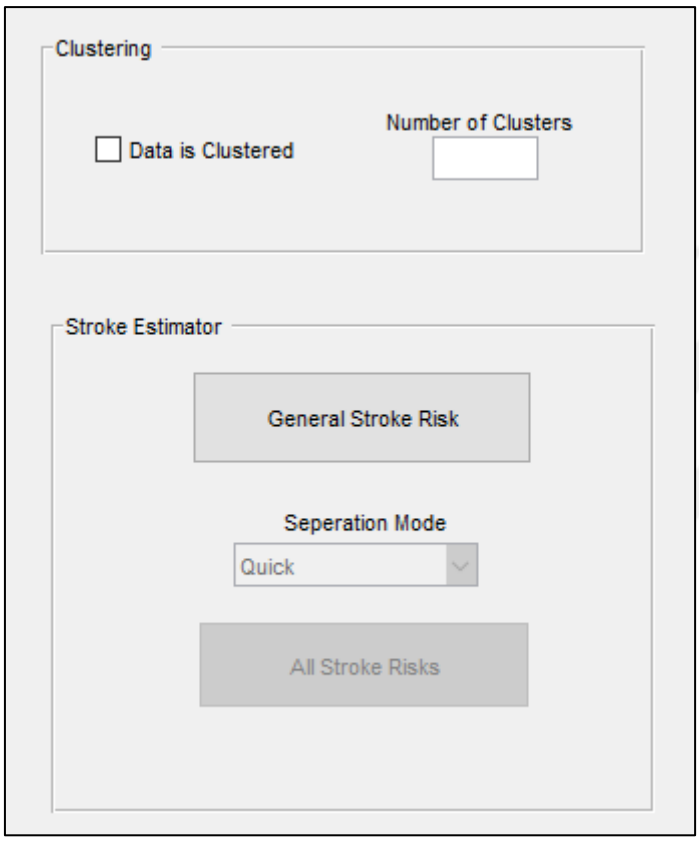


Figure 4.9. Clustering and stroke risk estimation control group

Clustering

The part of the GUI corresponding to clustering process consists of a checkbox and textbox with tags “Data is Clustered” and “Number of Clusters”, respectively. The checkbox is non-editable and it states whether the data is clustered or not at any point during operation. For instance, if the user selects a readily clustered dataset as training data, the checkbox will be marked, indicating that the data is clustered. In this case, the textbox “Number of Clusters” will display the number of clusters in the clustered dataset and becomes non-editable since it is not possible to change the number of clusters in a dataset that is already clustered. Otherwise, the textbox becomes editable after loading a training dataset in order that the user

can select an appropriate number of clusters for the clustering process. For the system to function properly, the number of clusters must be specified before initiating the stroke risk estimation. Upon clicking on “General Stroke Risk” pushbutton, the textbox becomes non-editable to prevent any modification in the dataset as the risk estimation process is started.

General stroke risk

The uppermost button of the stroke estimator group, “General Stroke Risk” pushbutton performs two important operations once pressed. Firstly, if the dataset is not readily clustered, the clustering module is called with number of clusters specified in the clustering textbox and clustering of the dataset is carried out accordingly. If the dataset is already clustered, this operation is skipped. After obtaining a clustered dataset, the risk estimation module is called to calculate the general stroke risk for the patients in the given dataset. By this button, the general stroke risk estimation process is separated from the IHTRE process for two main reasons. Firstly, the separation process is carried out using the results of the general stroke risk estimation, therefore the general stroke risk estimation results must be available by the time the user selects the separation mode. Secondly, in some cases, the user might be interested in only the general stroke risk in patients rather than individual stroke types. By this way, the user is given the option to run only the general stroke risk estimation part of the risk estimation module and avoid unnecessary computation that takes additional time.

After the general stroke risk is calculated for the given dataset, the popup menu “Separation Mode” becomes enabled. As its name suggests, this popup menu controls the separation mode with its options as “Quick” and “Full”. The selected option determines whether all samples or only the samples with nonzero general stroke risk will be transferred to IHTRE part of the risk estimation module as discussed in Section 4.1.

The last component in this group, “All Stroke Risks”, button corresponds to the separation and IHTRE parts of the risk estimation module. This button becomes enabled once the general stroke risk estimation process is completed. Once pressed, the selected option at “Separation Mode” menu is checked. If the “Quick” mode is chosen, the separation function is run and the samples with 0 estimated general stroke risk are separated from the dataset and their estimated risks of hemorrhagic stroke, ischemic stroke and TIA are set to 0 as well.

If the separation mode is chosen as “Full” no modifications are made on the dataset. After the separation process, the remaining (or all, if “Full” mode was selected) samples are sent to the IHTRE part of the risk estimation module and their risks of hemorrhagic stroke, ischemic stroke and TIA are determined as described in Section 4.2. This completes the operation of the proposed model with all 4 types of stroke risk estimation performed.

4.3.3. Risk estimation success

The success rate of proposed model for the current run is displayed in the four textboxes in terms of validation accuracy. Each of the four textboxes are activated after their associated stroke risk estimation is carried out. Therefore, the first textbox, “Stroke” displays the validation accuracy of the model in general stroke risk estimation after it is called using the “General Stroke Risk” pushbutton. The three remaining textboxes display the validation accuracy of their associated risk estimations as soon as the IHTRE process initiated by the “All Stroke Risks” button pressing is completed. A close-up view of the stroke estimation success panel after performing general stroke risk estimation is given in Figure 4.10. It is worth noting that the textboxes “Hemorrhage”, “Ischemia” and “TIA” are still blank since their associated risks are not estimated yet.

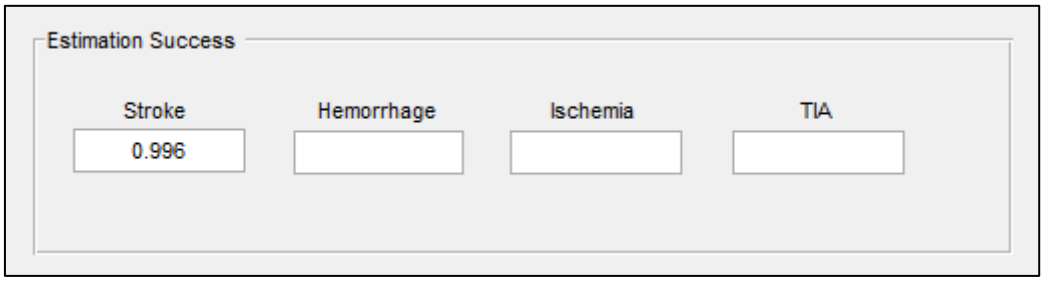


Figure 4.10. Estimation success part of the GUI, displaying validation accuracy values

4.3.4. Data viewer

The right side of the GUI is reserved for the components used to display the datasets and results used or obtained during the operation in various formats. This part is accommodated by two large table viewers, a popup menu and a pushbutton as illustrated in Figure 4.11. In this figure, the popup menu is pressed to display its available options.

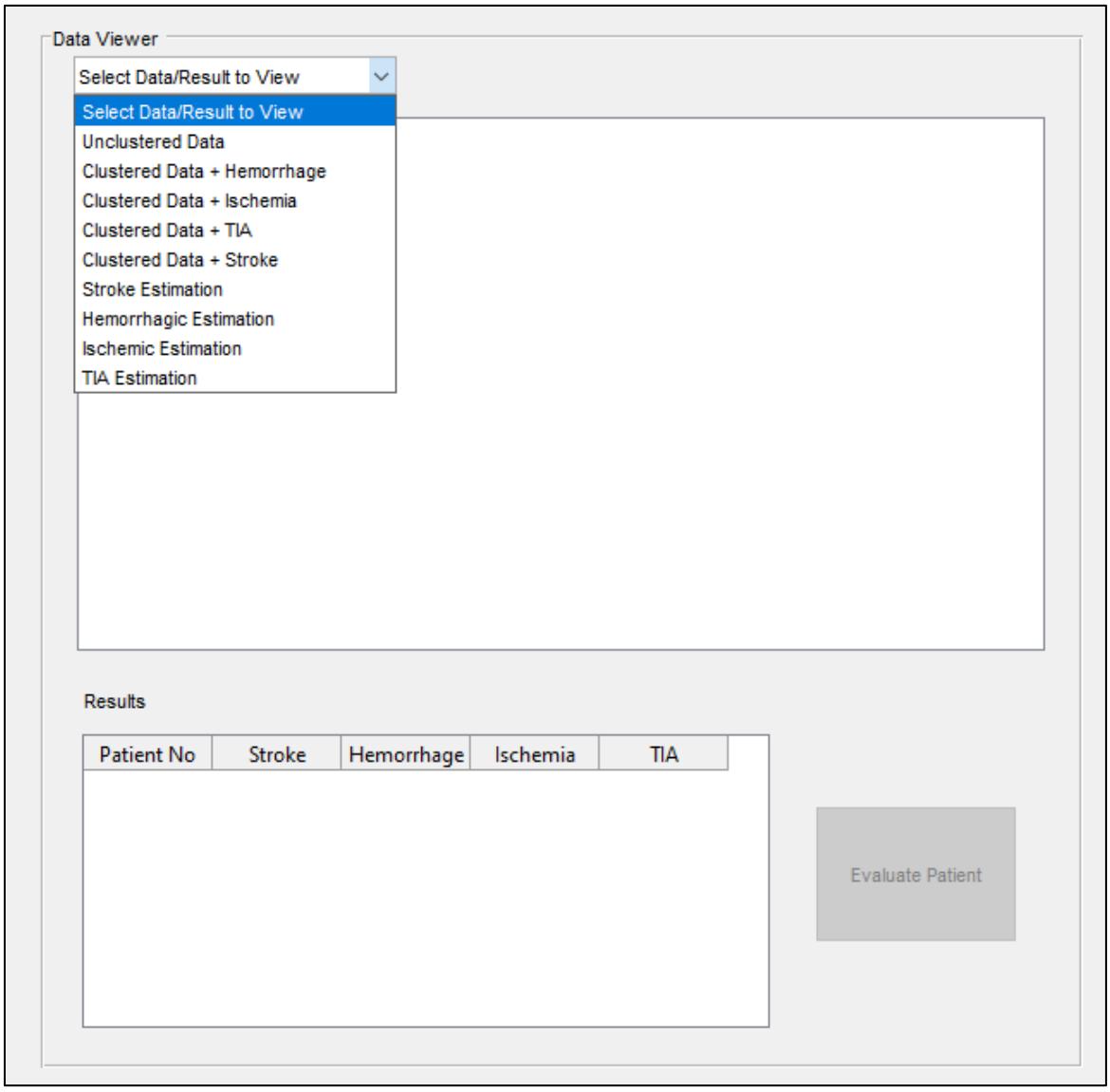


Figure 4.11. Data viewer part of the GUI with popup menu displaying available options

Configurable data and result viewer

The upper table viewer, paired by the popup menu constitutes the configurable data and result viewer that can be used to display a number of different elements in the GUI's workspace. The popup menu allows the user to select from 9 view options that determine what is displayed on the table viewer. The view options and their explanations are provided in Table 4.3. It should be noted that if the necessary data for the selected option is not generated or calculated yet, the table viewer will not display anything.

Table 4.3. View options and their explanations

Option	Explanation
Unclustered Data	Input dataset prior to clustering operation
Clustered Data + Hemorrhage	Clustered dataset with hemorrhagic stroke information as its last column.
Clustered Data + Ischemia	Clustered dataset with ischemic stroke information as its last column
Clustered Data + TIA	Clustered dataset with TIA information as its last column
Clustered Data + Stroke	Clustered dataset with general stroke information as its last column
Stroke Estimation	General stroke risk estimation results
Hemorrhagic Estimation	Hemorrhagic stroke risk estimation results
Ischemic Estimation	Ischemic stroke risk estimation results
TIA Estimation	TIA risk estimation results

Compact result viewer

While the configurable data and result viewer has the full functionality of displaying estimation results for all four stroke types, a second table viewer is employed to provide a more compact and all in one view of the estimation results. In this viewer, the risk estimation results of each of the four stroke types appear as soon as they are calculated by the risk estimation module. Unlike the configurable data and result viewer that displays the risk estimation and its binary version for the selected stroke type, the compact result viewer displays only the estimated risk for all patients and stroke types.

Patient evaluation

The GUI is complemented with a simple patient evaluation function that is designed primarily for non-medical users that might prefer an objective evaluation of the risk estimation results for a given patient. Coupled with the compact result viewer, the pushbutton “Evaluate Patient” generates a small dialog box displaying the evaluation of the selected patient in the table viewer according to his/her estimated risks. This evaluation provides information regarding the patient’s risk of having a stroke and the type of a possible stroke event. If the stroke risk of the patient is very low or there is no dominant type regarding risk estimates such as when risk estimations for multiple stroke types are considerably close to each other, the type of a possible stroke event is not stated.

5. RESULTS AND DISCUSSION

The proposed model discussed in Section 4 was given its final form according to a large number of analyzes carried out in the development stage alongside theoretical knowledge and expertise. The testing of the model during development as well as after it is completed is carried out by analyzes that are mostly focused on the performance of the model in terms of accuracy. The results of these tests are analyzed and used to optimize the proposed model during the development stage. After the model is given its final form, its performance is demonstrated and discussed to highlight its capabilities and imperfections.

The results are presented alongside relevant discussion in an order parallel to the operation of the proposed model, similar to the narration order of Section 4. All tests featured the dataset formed within this study and unless otherwise stated, all results presented in this section are obtained using 100 Monte Carlo trials.

5.1. Development Stage Results

In machine learning applications, the formal, theoretical knowledge of the subject is essential for creating a robust model and it provides guidance in determining the possible model parameters and features. However, unlike the majority of engineering branches, the optimization of a model cannot be carried out solely regarding theoretical information. For this reason, during the development of the proposed model, a number of performance analyzes are carried out to determine the optimal values of model parameters and features. These analyzes are performed in the form of controlled experiments during which only a single parameter was altered to observe its effect on the system performance.

In order to prevent settling for a local optimum, after determining the optimal value of a parameter, the tests carried out earlier were repeated with that optimal value to check if another combination yields better results. Therefore, the results presented in this section correspond to the tests that feature the optimal, and hence final values of all model parameters except the one associated with each test.

5.1.1. Clustering module results

The clustering module indirectly affects the model performance since it shapes the clustered dataset to be used in the risk estimation module. The tests regarding the clustering module are carried out to determine the optimal values of two main FCM parameters as well as which information provided by clustering process will be used in the risk estimation module.

Number of clusters

The optimal number of clusters for the proposed model is determined empirically, considering the accuracy achieved with different number of clusters for all four stroke types. The interval c=1:10 is scanned with c=1 corresponding to the case in which the dataset is directly input to the risk estimation module. For all other cases, the fuzzifier is set to m=1,56 and all information obtained as a result of clustering is included in the dataset provided to the risk estimation module. The results, expressed in terms of accuracy of the model in estimating four different stroke risks are given in Table 5.1. It should be noted that, the accuracy value corresponding to general stroke cases were taken as the main performance criterion.

Table 5.1. Accuracy (percent) of the model for different values of c

Number of Clusters (c)	General Stroke	Hemorrhagic Stroke	Ischemic Stroke	TIA
1	83,28	82,38	84,47	81,52
2	87,16	86,86	85,54	81,21
3	92,77	86,46	87,18	88,21
4	99,89	99,81	99,90	99,87
5	93,33	91,27	92,34	89,27
6	99,79	95,41	99,85	99,91
7	92,87	91,33	93,83	89,32
8	92,73	91,36	93,29	89,16
9	92,76	90,77	91,67	89,37
10	92,59	92,03	78,80	89,12

As can be seen from Table 5.1, for the base case the accuracy of the model is considerably lower than all other cases, indicating that clustering provides substantial improvement in the overall system performance. Moreover, the improvement is observed to be largest for general stroke risk estimation, exceeding 10% on average.

For c=2 it is observed that, the improvement provided by clustering process is relatively small. Moreover, the accuracy in estimating TIA is actually lower when compared to the base case, suggesting that clustering the patient dataset into two is not as beneficial, considering the increased workload introduced by the clustering module. However, for all other values of c, it is observed that the improvement in the system performance clearly overweighs the additional computational complexity due to the clustering module. The performance of the model for cases c=3, 5, 7, 8, 9 and 10 exhibited similar behavior with the exception of ischemic stroke estimation for c=10. Furthermore, the cases c=4 and c=6 provide near-perfect results with accuracy values approaching 100% in almost all stroke types. However, c=6 case provided a slightly lower accuracy in estimating the risk of hemorrhagic stroke when compared to c=4 case alongside marginally lower accuracies in general stroke risk and ischemic stroke risk. For these reasons, the optimal number of clusters is determined as c=4 for the proposed model with c=6 being an almost equal alternative. These results can be observed in a single graph as given in Figure 5.1 for a visual comparison, highlighting the two optimal peaks corresponding to c=4 and 6 for each stroke type, surrounded by lower-accuracy regions.

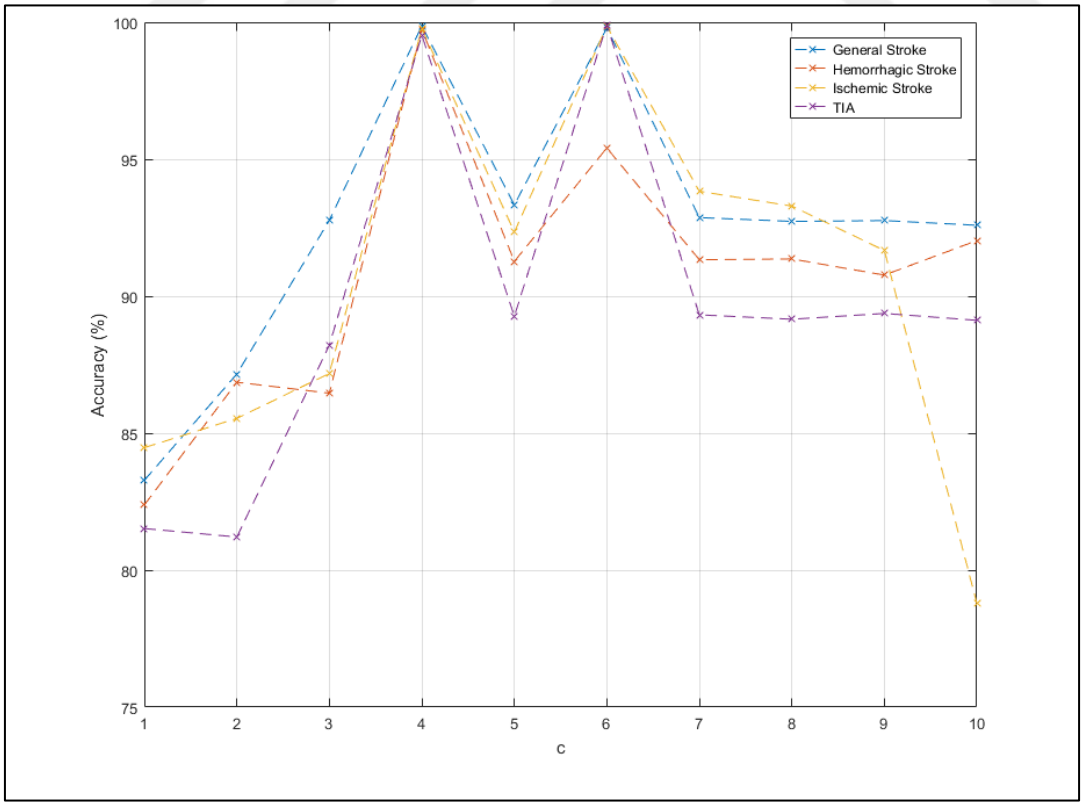


Figure 5.1. Graphical visualization of the accuracy versus c results given in Table 5.1.

Fuzzifier

The fuzzifier expressed by m, is one of the most important parameters of FCM, determining the fuzziness of the clustering process. The optimal value for fuzzifier is determined using an exhaustive search, scanning the region m ∈ [1 5] with a step size of 0,01. For each value of m, the accuracy of the model is calculated for all stroke types using 100 Monte Carlo trials. The accuracy in estimation of general stroke risk is selected as the main performance criterion in the case of the results regarding other stroke types breaking approximately even for multiple values of m. The graphs demonstrating the accuracy of model with respect to m values for each of the four stroke types are given in Figures 5.2 through 5.5.

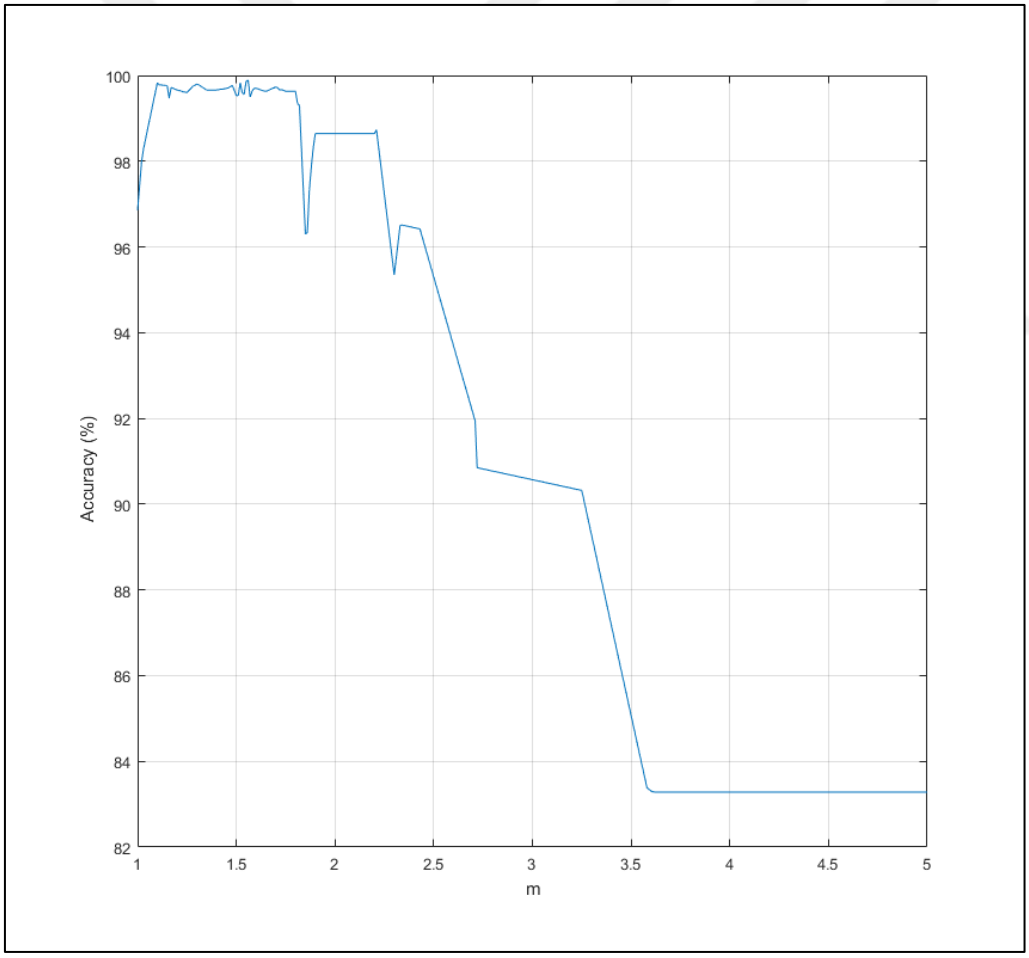


Figure 5.2. Accuracy of the model in estimating general stroke risk with respect to m

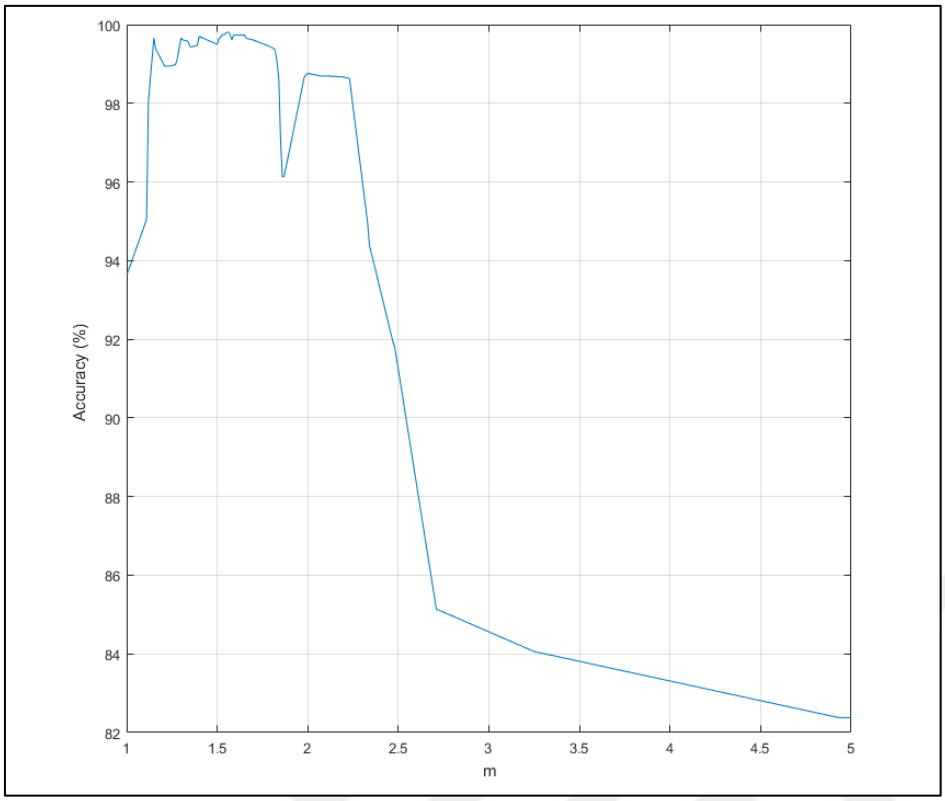


Figure 5.3. Accuracy of the model in estimating hemorrhagic stroke risk with respect to m

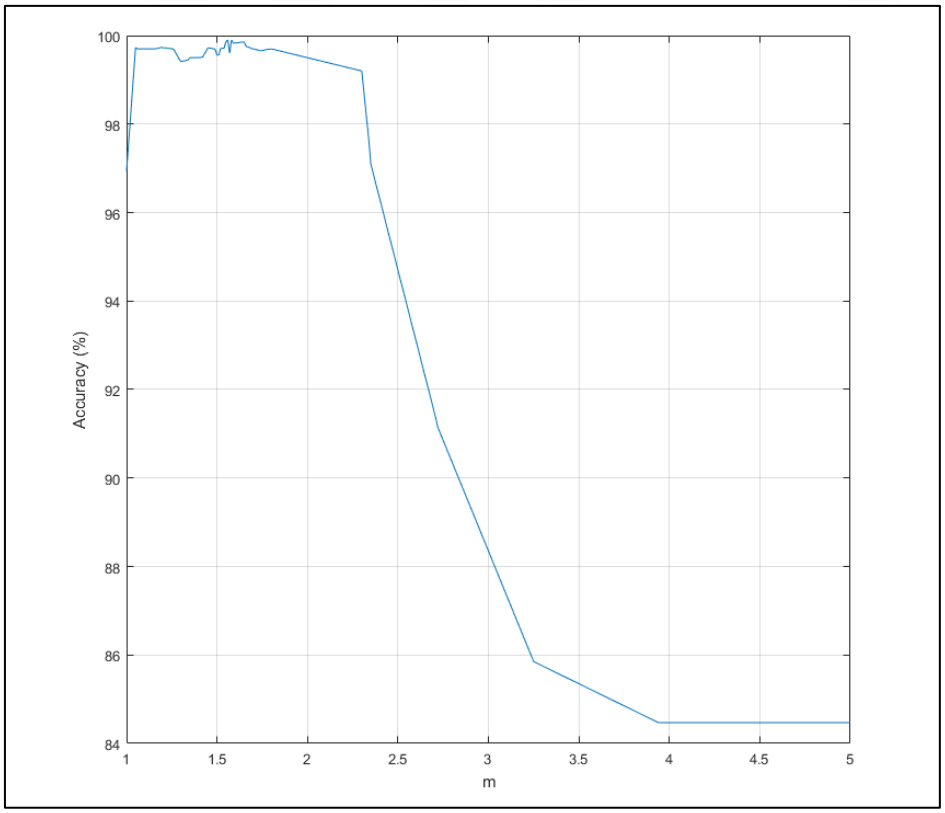


Figure 5.4. Accuracy of the model in estimating ischemic stroke risk with respect to m

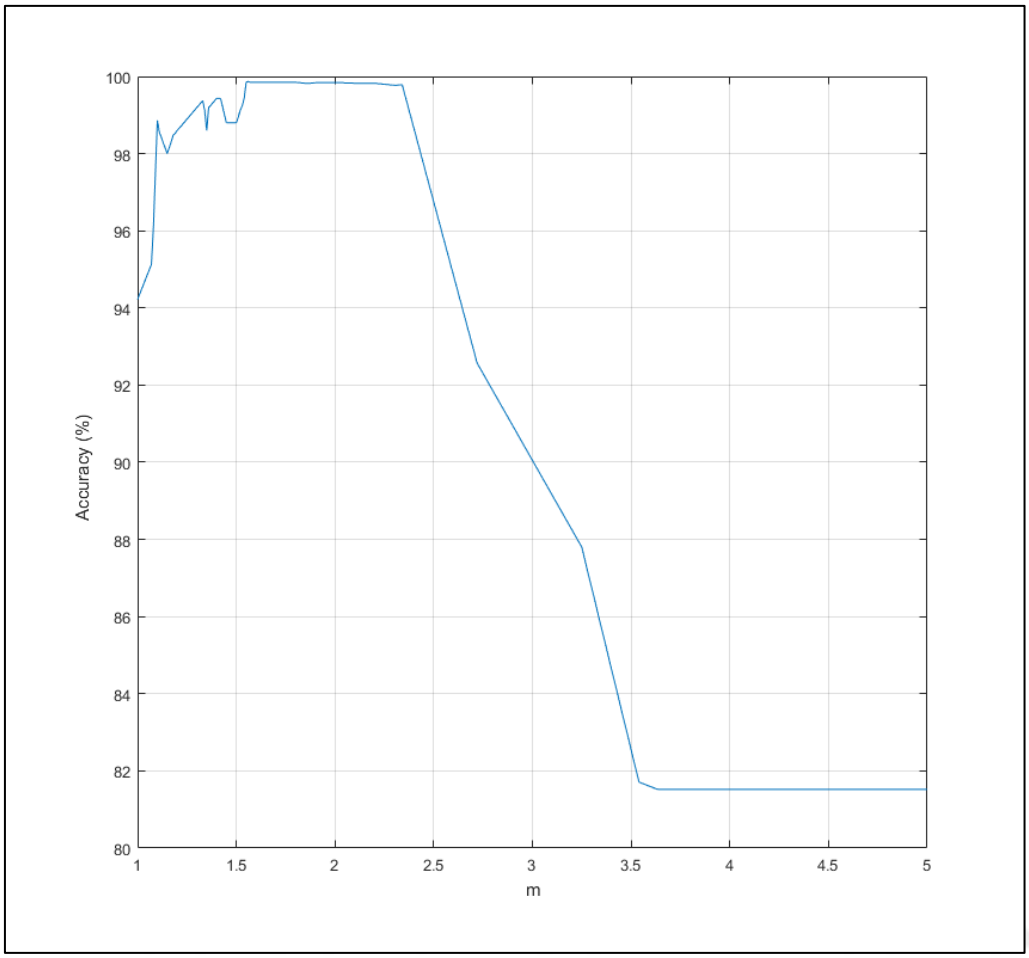


Figure 5.5. Accuracy of the model in estimating TIA risk with respect to m

For all stroke types, as demonstrated in Figures 5.2 to 5.5, it is observed that for m=1 which corresponds to crisp clustering that is similar to k-means clustering, the accuracy of the model starts at fair values and increases rapidly with m. This brief yet sharp increase saturates with the accuracy approaching 100% for all stroke types which is observed as plateaus starting not far from m=1. With m increasing beyond 2, the accuracy for all models start to decrease sharply with the exact point and rate of decrease depending on the stroke type. For large values of m, it is observed that the accuracy saturates at its lower boundary that is no different than the base (c=1) case demonstrated in Table 5.1.

The decrease in accuracy for large values of m can be explained by the fact that, as m is increased beyond some limit, the clustering process becomes overly fuzzy, causing the clusters lose their discriminative properties. This results in the membership grades becoming arbitrarily close to each other for almost all samples in the dataset, hence the expressive power of clusters disappears. This phenomenon can be observed in the fragment of clustered

dataset obtained for m=3,50, as given in Figure 5.6. In this figure, columns 125 to 128 represent the membership grades of given samples for clusters 1 to 4, respectively. Column 129 displays the maxc values for each sample.

125	126	127	128	129
0.2438	0.2438	0.2686	0.2438	3
0.2333	0.2333	0.3000	0.2333	3
0.2568	0.2568	0.2297	0.2568	2
0.2393	0.2393	0.2821	0.2393	3
0.2381	0.2381	0.2857	0.2381	3
0.2249	0.2249	0.3252	0.2249	3
0.2491	0.2491	0.2527	0.2491	3
0.2569	0.2569	0.2294	0.2569	2
0.2826	0.2827	0.1520	0.2827	2
0.2382	0.2382	0.2854	0.2382	3
0.2423	0.2423	0.2731	0.2423	3
0.2468	0.2468	0.2597	0.2468	3
0.2320	0.2320	0.3040	0.2320	3
0.2505	0.2505	0.2485	0.2505	2
0.2574	0.2574	0.2277	0.2574	1
0.2725	0.2725	0.1824	0.2725	2
0.2647	0.2648	0.2057	0.2648	2
0.2557	0.2558	0.2327	0.2558	2
0.2813	0.2813	0.1562	0.2813	2
0.2597	0.2597	0.2210	0.2597	2
0.2412	0.2412	0.2763	0.2412	3
0.2368	0.2368	0.2895	0.2368	3
0.2561	0.2561	0.2316	0.2561	2
0.2763	0.2763	0.1712	0.2763	2

Figure 5.6. Membership grades and maxc of a fraction of samples for m=3,50

Therefore, it can be concluded that FCM clustering provides substantial improvement in model accuracy only with a properly selected value of m. Otherwise, a crisp clustering algorithm such as k-means clustering would be much useful than an overly fuzzy FCM clustering algorithm.

When Figure 5.2 is analyzed, it is observed that the estimation accuracy of general stroke risk exhibits a rougher trend when compared to other stroke types with multiple step-like descents are observed alongside two local minima. Furthermore, the accuracy reaches 99% as early as m=1,07 exhibiting a rather steep increase. The accuracy remains above this

threshold until m exceeds 1,82 with maxima occurring at m=1,56. After this value, a second, smaller plateau corresponding to the vicinity of 96% accuracy is observed that persists until around m=2,40. With increasing value of m, the accuracy drops sharply with a mild region observed between m=2,70 and m=3,20. After this point the accuracy of the model decreases rapidly, reaching the global minima of 83,28% at m= 3,62. After this point, the accuracy is invariant to increase in fuzzifier, indicating that the overly fuzzy clustering could not provide the model any useful information regarding general stroke risk estimation.

Figure 5.3 demonstrates the variation of model accuracy in hemorrhagic stroke risk estimation with respect to m. For this case, a relatively rough plateau starts at m=1,14 with a damped oscillational behavior that settles down at around m=1,70. The maximum accuracy occurs within this region at m=1,56. A sharp local minima is observed at m=1,86 that paves the way to the second, smaller plateau persisting until m=2,30. After that value of m, the accuracy of the model decreases sharply, in a similar manner to all other stroke types. However, the accuracy reaches global minima much later than other types, at around m=4,92 implying that the model was able to utilize what little guidance the overly fuzzy classification process provided.

In Figure 5.4, the accuracy of the model in estimating ischemic stroke risk is observed with respect to the fuzzifier value. Among all stroke types, ischemic stroke is the case for which the accuracy increases most rapidly after m=1. Furthermore, it remains fairly stable until around m=2,25 where a steep and steady descent starts. The maximum accuracy is observed at multiple values of m for this case, including m=1,56. Unlike general and hemorrhagic stroke risk, there are no sharp local minima or second plateau phenomenon indicating that the ischemic risk estimation process is relatively invariant to small changes in the fuzzifier value. The steep descent leaves its place to a shallower one at around m=3,22 that ends at m=3,94 with the accuracy reaching the global minima.

The accuracy results for TIA risk estimation with respect to m is given in Figure 5.5. The initial increase in accuracy is relatively shallower and exhibits a slightly more unstable behavior for this case when compared to other stroke types. However, when the accuracy is settled at around m=1,52 it remains fairly constant until m approaches 2,30 indicating that the model performance is optimal within this region. The exact maxima occurs at m=1,56 corresponding to an accuracy of 99,87%. The steep decrease starting at m=2,30 is similar to

what is observed for the ischemic stroke case. The saturation of accuracy at global minimum occurs at around m=3,64 which is relatively early when compared to ischemic and hemorrhagic stroke cases and approximately same as the general stroke case.

Considering the results obtained for all stroke types, the optimal value of m is determined to be 1,56 for the final model. While this value of fuzzifier yields highest accuracy in estimating all four stroke types, the flat regions around this value as observed in Figures 5.2 to 5.5 could also be considered as the possible range of m for similar applications.

Utilization of clustering outputs

As discussed in Section 4.1.8, the FCM algorithm provides the membership grades of each sample as the default output. Another output, expressed as maxc, is obtained by taking the index of highest membership grade associated with each sample which corresponds to the cluster that a given sample belongs to most. For 4 clusters, there are 5 outputs that could be possibly used to modify the dataset, 4 corresponding to the membership grades and 1 corresponding to the maxc value of each sample. These outputs could then be merged with the input dataset to form the clustered dataset that will be transferred to the risk estimation module. In order to determine the effect of including the membership grades and maxc values on the performance of the model, four different cases depending on which outputs of the clustering algorithm are included in the clustered dataset are considered:

- Base Case: No clustering performed
- Membership Grade Case: Only the 4 membership grades are included in the dataset
- Maxc Case: Only the maxc feature is included in the dataset
- Full Case: All of the clustering algorithm’s outputs are included in the dataset

The performance of the system in estimating four types of stroke is analyzed for each case with 100 Monte Carlo trials. The result of these analyzes, expressed in terms of accuracy of the model is given in Table 5.2. The highest accuracy associated with each stroke type is highlighted in bold.

Table 5.2. The accuracy of the model in estimating four types of stroke for different cases

Case	General Stroke	Hemorrhagic Stroke	Ischemic Stroke	TIA
Base	83,28	82,38	84,47	81,52
Membership Grade	98,69	99,28	97,95	95,75
Maxc	96,84	93,63	96,93	94,22
Full	99,89	99,81	99,90	99,87

When Table 5.2 is analyzed, it is observed that the base case has the lowest accuracy for all stroke types as expected. Including either the membership grade or maxc information in the clustered dataset is shown to increase the model accuracy substantially for all stroke types. Furthermore, the improvement in accuracy provided by membership grade inclusion turned out to be higher than that provided by maxc inclusion.

Inclusion of membership grades provided its largest improvement in hemorrhagic stroke risk estimation task with an absolute increase in accuracy by 16,90% while its smallest improvement is observed for TIA risk estimation where it provided an increase of 14,23%. Inclusion of maxc values, on the other hand, was shown to be most effective for general stroke risk estimation, providing an absolute in accuracy by 13,56% which is still lower than the smallest improvement inclusion of membership grades provided. The smallest improvement in this performance analysis with respect to the base case corresponded to inclusion of maxc values for hemorrhagic stroke risk estimation with an increase of 11,25%. The difference in individual improvements is highest for hemorrhagic stroke case for which inclusion of membership grades provided an absolute increase in accuracy that is 5,65% higher than what is provided by inclusion of maxc values.

On the other hand, including membership grades alongside maxc values in the clustered dataset resulted in highest performance for all stroke types. The improvement is observed to be largest for TIA risk estimation with an absolute increase in accuracy by 18,35%. The smallest improvement in performance for the full case is achieved in ischemic stroke risk estimation with 15,43% absolute increase in accuracy. It should be noted that the difference between these improvements is largely due to the difference in the accuracy of base model for different stroke types while that between accuracies of the full case is considerably small.

Therefore, it can be concluded that the information conveyed by membership grades and maxc features contribute to the classification process independently, acting as uncorrelated feature groups. In the light of these results, the final model is determined to employ membership grades and maxc features together to form the clustered dataset.

5.1.2. Risk estimation module results

The risk estimation module employs 16 different base models to perform risk estimation of 4 different types of stroke as described in Section 4.2. The output of this module, which is also the output of the whole model, is determined according to the individual outputs of these base models employing dynamically adjusted weights. Therefore, it is suggested that the performance of the final model does depend on the performance of individual base models as well as the weights that determine their contribution on the final output. The tests associated with the risk estimation module are carried out in order to demonstrate the individual performance of the base models and the advantage of utilizing dynamic weights.

Performance of the base models

The performance of the base models used to form the classifier ensembles are analyzed considering their accuracy and AUC values. The results of these analyzes are provided in Tables 5.3 through 5.6, corresponding to general stroke, hemorrhagic stroke, ischemic stroke and TIA risk estimation cases, respectively. The column “Difference” in these tables represent the difference between the accuracies of the given base model and proposed model in estimating the risk of the related stroke type.

Table 5.3. Performance of base models in general stroke risk estimation

Base Model	Accuracy (%)	AUC	Difference
Standard Tree	88,44	0,94	11,45
Boosted Tree	92,03	0,92	7,86
Bagged Tree	89,13	0,92	10,76
Balanced Gaussian SVM	77,62	0,91	22,27

Table 5.4. Performance of base models in hemorrhagic stroke risk estimation

Base Model	Accuracy (%)	AUC	Difference
Standard Tree	85,72	0,92	14,09
Cosine kNN	95,89	0,95	3,92
Weighted kNN	92,17	0,93	7,64
Linear SVM	92,03	0,95	7,78

Table 5.5. Performance of base models in ischemic stroke risk estimation

Base Model	Accuracy (%)	AUC	Difference
Standard Tree	88,42	0,89	11,48
Fine kNN	90,92	0,95	8,98
Cosine kNN	89,22	0,92	10,68
Boosted Tree	90,05	0,96	9,85

Table 5.6. Performance of base models in TIA risk estimation

Base Model	Accuracy (%)	AUC	Difference
Simple Tree	85,43	0,90	14,44
Fine Gaussian SVM	95,92	0,94	3,95
Minkowski kNN	89,59	0,92	10,28
Balanced kNN	84,96	0,89	14,91

When Tables 5.3 to 5.6 are analyzed, it is observed that the accuracy of the base models vary between 77,62% and 95,92%. On the other hand, the AUC values are noted to be confined to a relatively narrower range, varying between 0,89 and 0,96.

Table 5.3 provides the Accuracy and AUC values for the base models used in general stroke risk estimation task. For this case, the boosted tree model performed better than others with an accuracy of 92,03% while the AUC value was highest for the standard tree model. Balanced Gaussian SVM, being the only base model that is not based on decision trees for general stroke estimation, performed relatively poorly, achieving 77,62% accuracy, which is the worst accuracy among all 16 base models. It should be kept in mind that, the base models used in the final model was selected among the best performing ones from a list of 25 different models for each stroke type. Therefore, while the Balanced Gaussian SVM model seems to perform poorly, it still performed better than the remaining 21 models tested

during the formation of risk estimation module. For general stroke risk estimation task, it can be claimed that tree-based models turned out to be dominant by means of performance and hence they are considered as most suitable base models for this case.

The performance of base models employed in hemorrhagic stroke risk estimation can be seen in Table 5.4. For this case, it is observed that the two kNN models performed better than the remaining base models with Cosine kNN achieving 95,89% accuracy, ranking second among all base models. Furthermore, the difference between the accuracy of Cosine kNN and final model was only 3,92% which is the lowest difference recorded for all 16 base models. The AUC value for Cosine kNN was also the highest, although in par with Linear SVM which achieved a lower accuracy. Among the base models considered for this case, Standard Tree model performed with least accuracy alongside the lowest AUC value.

Table 5.5 demonstrates the results associated with the base models in ischemic stroke risk estimation. It is observed that, the accuracies of the base models are noticeably close to each other, unlike the other cases. Among these base models, Fine kNN performed slightly better with an accuracy of 90,92% while the highest AUC value is recorded for Boosted Tree model. The Standard Tree model, on the other hand, performed relatively poorly with lowest accuracy and AUC values among the four base models.

The accuracy and AUC values obtained for base models in TIA risk estimation task is given in Table 5.6. For this case, Fine Gaussian SVM model is observed to perform significantly better than the remaining base models, achieving 95,92% accuracy, which is the best accuracy among all 16 base models. This model also had the highest AUC value among the base models employed for TIA risk estimation task. Balanced kNN model, on the other hand, was the model with lowest accuracy and AUC value, performing slightly worse than Simple Tree model.

The result of these analyzes showed that, for all stroke types, the ensembles formed using the base models performed substantially better than any base model used individually. Moreover, the improvement provided by ensemble formation was observed to be above 7,5% except for two cases, namely Cosine kNN for hemorrhagic stroke and Fine Gaussian SVM for TIA risk estimation. Therefore, it can be concluded that the four classifier

ensembles, formed using four base models for each ensemble, provide a considerable improvement in overall model performance.

Effect of adaptive weights

The effect of using adaptive weights in the risk estimation module on model performance can be demonstrated by a controlled experiment that compares the performance of a model that employs fixed and equal weights to that of proposed model. For this model, called fixed weight model from now on, the expression of weights calculated in Eq. 4.6 is modified as given in Eq. 5.1.

$$w_{ik} = 1 \tag{5.1}$$

where w_{ik} is the weight and associated with i^{th} base model and k^{th} cluster. This resulted in each base model having an equal contribution in the decision making process associated with the risk estimation task for 4 stroke types. The performance of the fixed weight model, in terms of accuracy, is given in Table 5.7 alongside the performance of the final model as well as the best performing (in terms of accuracy) base model for each stroke type. The results for the base case described in Table 5.2 was also included to compare the effect of including clustering outputs with that of adaptive weights.

Table 5.7. Performance of the fixed weight model, final model and base models

Model	General Stroke	Hemorrhagic Stroke	Ischemic Stroke	TIA
Fixed Weight	98,05	97,53	92,78	97,67
Proposed	99,89	99,81	99,90	99,87
Best Performing Base	92,03	95,89	90,92	95,92
Base (c=1)	83,28	82,38	84,47	81,52

When Table 5.7 is analyzed it is noted that, while performing worse than the proposed model, the fixed weight model performed better than the best performing base model for all stroke types. The accuracy of the fixed weight model was maximum in estimating the general stroke risk while it performed relatively poorly in ischemic stroke risk estimation task which was the one that the proposed model performed best. Therefore, the difference in performances of fixed weight and proposed models is observed to be maximum for ischemic

stroke case with an absolute difference of accuracy of 7,12%. On the other hand, for general stroke risk, both models performed close to each other, with fixed weight model achieving 1,84% lower accuracy than the proposed model.

When the fixed weight model is compared with the best performing base models, it is observed that the difference of accuracy between these models is observed to be maximum for general stroke case with 6,02% while it turned out to be minimum for hemorrhagic stroke risk estimation task with only 1,64%. Furthermore, it is noted that the average difference in accuracy values of fixed weight and best performing base models are slightly lower than that between fixed weight and proposed models. From this perspective, it can be claimed that introduction of dynamic weights has provided a larger improvement on overall performance when compared to formation of classifier ensembles. On the other hand, the accuracy improvement provided by including the clustering results in the dataset is observed to be largest. Taking everything into account, it is made clear that usage of dynamically adjusted weights in the risk estimation module has contributed to the model performance considerably, resulting in a robust and accurate model operation.

5.2. Final Results

After the proposed model is given its final form according to the results presented in Section 5.1, its performance is assessed through a number of analyzes considering a variety of performance metrics. The results presented in this section, as the ones presented earlier, are obtained with the dataset formed within the context of this study. Furthermore, the performance of the model is analyzed considering different separation ratios for the main dataset to observe the effect of training dataset size on the results. Results of some experiments that are not directly related to the correctness of model outputs such as operation time or memory requirement are also provided at the end of this section as auxiliary results.

5.2.1. Performance analysis results

The performance of the final model is analyzed using different metrics to improve the comparability of the proposed model with similar studies in literature, both past and future. By this way, it is aimed to provide improved validity of the results obtained in this study objectively.

Accuracy

Accuracy is the principal performance criterion used in this study to evaluate the correctness of the results provided by the proposed model. The accuracy of the model is considered regarding its four different outputs, representing risk estimates for four different stroke types. Therefore, it would be more relevant to talk about accuracy associated with each stroke type rather than a single accuracy value. The accuracy of the proposed model is obtained as 99,89% in general stroke risk estimation, 99,81% in hemorrhagic stroke risk estimation, 99,90% in ischemic stroke risk estimation and 99,87% in TIA risk estimation tasks. Considering these values, it can be claimed that the overall performance of the proposed model is exceptionally high with small variations in accuracy observed among different stroke types. Among the model outputs, it is observed that the highest accuracy is observed for ischemic stroke risk estimation while the lowest accuracy is obtained for hemorrhagic stroke risk estimation.

Accuracy with smaller training datasets

The accuracy values presented earlier are originally obtained by separating the dataset as training and test data with a ratio of 9:1 as mentioned in Section 4.1.8. This corresponded to the training dataset containing 2700 samples while the test dataset consisted of 300 samples. In order to analyze the effect of using a smaller training dataset, as well as reserving more samples to the test dataset, a number of tests involving training and test datasets of different size are carried out. It should be noted that, since the separation ratio is different than 9:1, the number of folds in the cross validation is no longer 10 for these datasets. The results for these datasets, alongside the original results discussed earlier, are provided in the form of estimation accuracy for each stroke type in Table 5.8.

Table 5.8. Accuracy of the model for the cases with different dataset separation ratios.

Separation Ratio	General Stroke	Hemorrhagic Stroke	Ischemic Stroke	TIA
9:1	99,89	99,81	99,90	99,87
4:1	99,63	99,79	99,70	99,87
7:3	99,60	99,78	99,67	99,86
3:2	99,56	99,78	99,66	99,86
1:1	99,50	99,73	99,60	99,86

As can be seen from Table 5.8, the accuracy of the model shows little deterioration with shrinking training dataset indicated by the decrease in separation ratio. The model performance is observed to be relatively sensitive to dataset size for general and ischemic stroke risk estimation since the decrease in accuracy was more noticeable for these cases. On the other hand, TIA risk estimation process turned out to be almost completely insensitive to changes in training dataset size, with only a 0,01% decrease observed in model accuracy between the two extremes of training set size.

The smallest training dataset analyzed consisted of 1500 samples, corresponding to a 1:1 separation ratio with 2-fold cross validation. For this case, the model achieved highest accuracy in TIA risk estimation task with 99,86% while the lowest accuracy of the model was recorded as 99,50% in general stroke risk estimation. Taking the results presented in this section into account, the system is proven to be substantially robust with respect to training dataset size, even when considerably small training sets are employed.

Confusion matrix

Confusion matrix is a simple yet powerful tool that provides essential information about the correctness of the predictions carried out by a model [188]. A number of different performance metrics, including accuracy and precision, could be obtained from simple operations carried out on the elements of a confusion matrix.

The results presented in the confusion matrix formed for this study is obtained by performing 10-fold cross validation on the dataset with 9:1 separation ratio. In order to evaluate the performance of the model over all samples in the dataset, 10 different training-test dataset pairs are formed, samples of which are selected randomly without replacement. This process made it possible to utilize all 3000 samples as test data. The results regarding these samples are demonstrated via 4 different confusion matrices corresponding to the model performance in estimating the risk of 4 types of stroke.

The confusion matrices obtained in the study are provided in Figures 5.7 through 5.10, corresponding to general stroke, hemorrhagic stroke, ischemic stroke and TIA risk estimation cases, respectively. The elements of the confusion matrices shown in these figures

are obtained by 10 Monte Carlo trials for each of the 10 training-test dataset pairs, corresponding to 100 trials per matrix and are rounded to the nearest integer value.

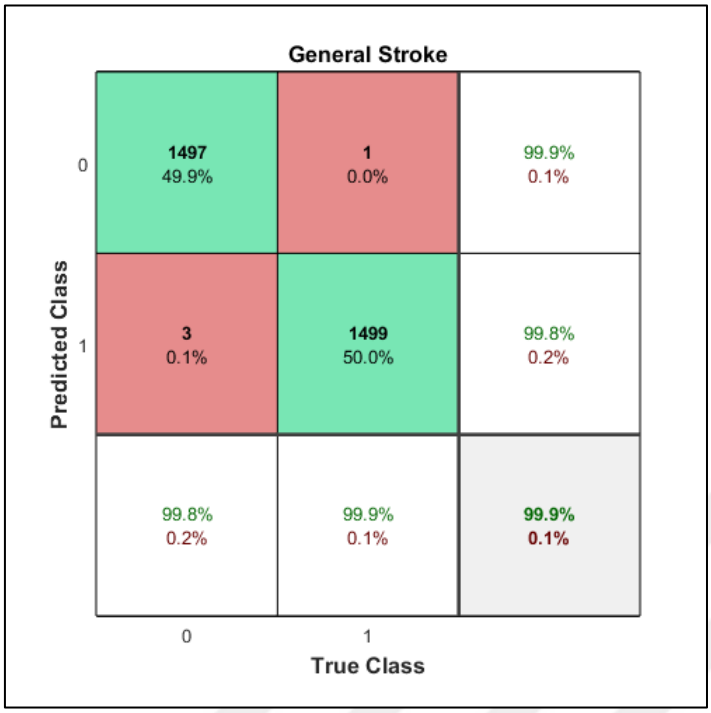


Figure 5.7. Confusion matrix for general stroke risk estimation

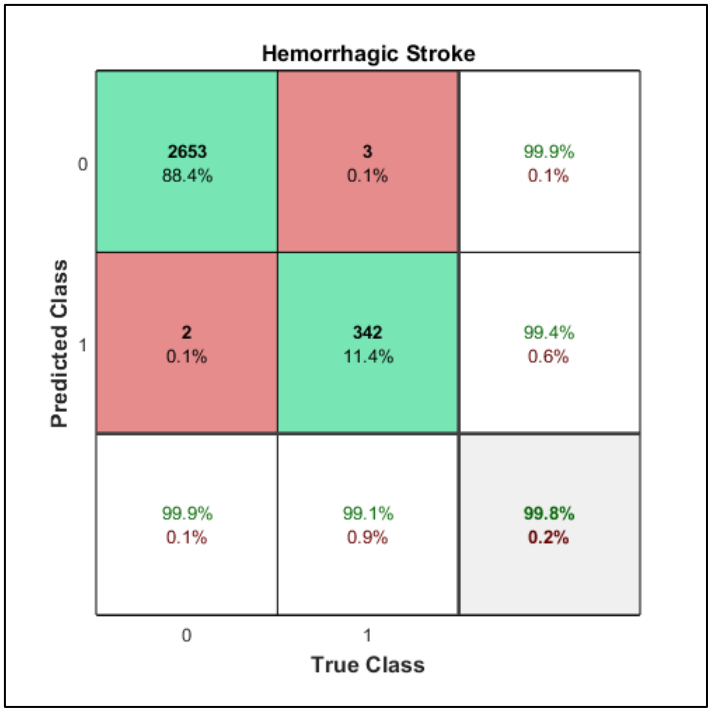


Figure 5.8. Confusion matrix for hemorrhagic stroke risk estimation

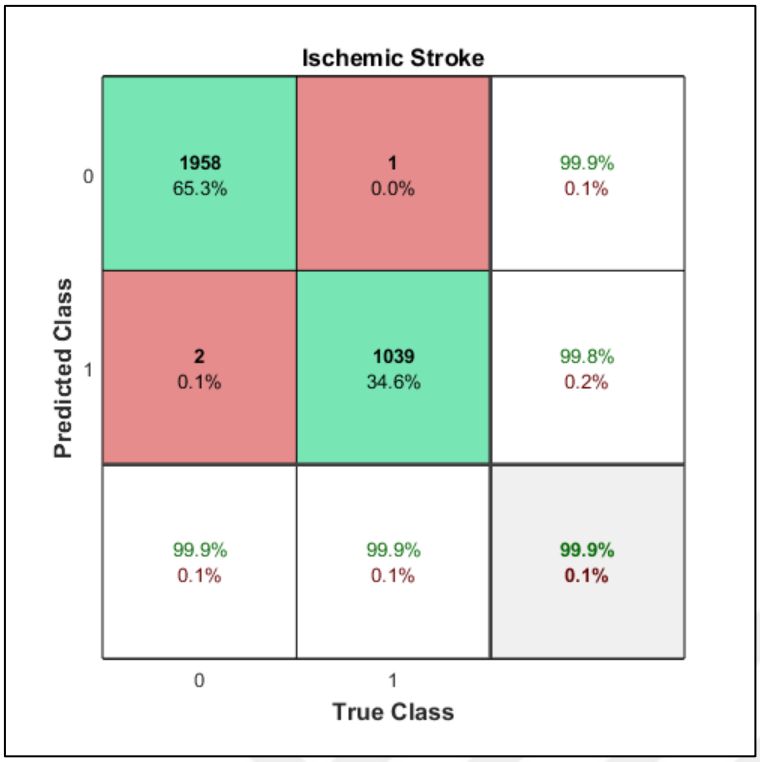


Figure 5.9. Confusion matrix for ischemic stroke risk estimation

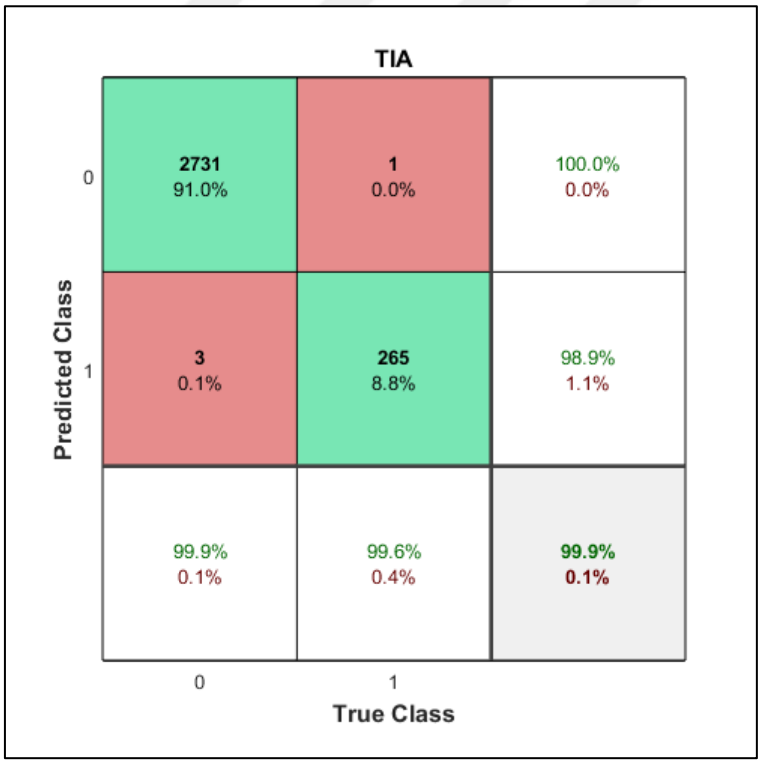


Figure 5.10. Confusion matrix for TIA risk estimation

For all stroke types, the model is observed to achieve very high degrees of correctness, which is an expected result considering the model accuracy values obtained earlier which can also be verified from the lower right corner of each matrix. Using these confusion matrices, four alternative identifiers of correctness, namely; true positive, false positive, true negative and false negative rates can be calculated as given in Table 5.9.

Table 5.9. Correctness rates (in per cent) for the results of stroke risk estimation tasks

Rate	General Stroke	Hemorrhagic Stroke	Ischemic Stroke	TIA
True Positive	99,93	99,13	99,90	99,62
False Positive	0,20	0,58	0,19	1,12
True Negative	99,80	99,92	99,90	99,89
False Negative	0,06	0,11	0,05	0,04

As can be seen from Table 5.9, for all stroke types the true positive and true negative rates are considerably close to 100%. The highest true positive rate is obtained for general stroke case while the highest true negative rate is obtained for the hemorrhagic stroke case. TIA risk estimation task is noted to be the one with highest false positive rate while hemorrhagic stroke case has the highest false negative rate. Furthermore, the false positive rates are observed to be higher for all stroke types when compared to false negative rates. This implies that the proposed model tends to overestimate the stroke risk, which is considered to be more desirable than underestimating the risk of such a serious health condition.

F-score

Although not formally selected as the performance criterion for this study, variants of F-score can also be easily calculated from the information provided by Figures 5.7 to 5.10. The expressions for the F1 and F2 scores, which are the two variants frequently used in the literature; are given in Equations 5.2 and 5.3, respectively [189].

$$F1 = \frac{TP}{TP+0,5(FP+FN)} \tag{5.2}$$

$$F2 = \frac{5 \cdot TP}{5 \cdot TP+4 \cdot FN+FP} \tag{5.3}$$

where TP, FP, FN and FP are the number of true positive, false positive, false negative and false positive observations, respectively. The resulting F1 and F2 scores of the model, obtained separately for each stroke risk estimation task is provided in Table 5.10.

Table 5.10. F1 and F2 scores of the proposed model in four stroke risk estimation tasks

Score	General Stroke	Hemorrhagic Stroke	Ischemic Stroke	TIA
F1	0,999	0,993	0,999	0,993
F2	0,999	0,992	0,999	0,995

As Table 5.10 suggests, both F1 and F2 scores are considerably close to unity which corresponds to perfect classification. For general and ischemic stroke risk estimation tasks, resulting F1 and F2 scores are equal at 0,999; the highest scores obtained in this study. On the other hand, F1 and F2 scores associated with hemorrhagic and ischemic stroke risk estimation cases recorded to be slightly lower. F2 score, which weighs recall higher than precision, was observed to be lowest for hemorrhagic stroke risk estimation case.

5.2.2. Auxiliary results

There are a number of measures that could be used when evaluating the performance of a model that are not directly related to the correctness of the results but are still important as they provide information about characteristics of the model in other ways. The results regarding the operation time and memory requirement of the proposed model are presented in this section as auxiliary results.

Operation Time

The operation time of the proposed model is measured using the “tic” and “toc” time marker commands of Matlab at relevant points of the code. Time required for completion of clustering, general stroke risk estimation and IHTRE processes are recorded individually. Each result is averaged over 100 Monte Carlo trials and the (full) completion time is obtained by summation of the results for clustering, general stroke risk estimation and IHTRE processes to filter the duration of user interaction such as typing a value or choosing a folder.

The operation time for each process, as well as the completion time of the model are given in Table 5.11. It should be noted that these results are obtained on an Intel Core i7 2,2 GHz machine running Ms Windows 10 Home operating system with Matlab R2016a software.

Table 5.11. Operation time results for the proposed model, in seconds

Process	Clustering	General Stroke Risk Estimation	IHTRE	Completion Time
Duration (s)	1,51	1,87	5,17	8,56

When Table 5.11 is analyzed, it is observed that among the three main processes of the proposed model, the clustering process takes the least amount of time while the IHTRE process requires the highest amount of time to complete. The IHTRE process, which includes estimation of three types of stroke risk, takes 2,76 times longer than the general stroke risk estimation process, which involves only one risk estimation task instead of 3. This implies that a noticeable fraction of time is spent on auxiliary operations such as accuracy calculation and GUI display that are carried out once for both processes. The actual duration of each risk estimation process alongside that of auxiliary operations can be estimated with the assumption that risk estimation for each stroke type takes an equal amount of time. This assumption is a valid one since the risk estimation is carried out employing the same algorithm for each stroke type with only differences being in the inputs and classifiers. With this assumption, the estimation of duration of a single risk estimation process reduces to solving a simple equation pair as given in Eq. 5.4

$$T_E + T_A = 1,87 \; ; \; 3T_E + T_A = 5,17 \tag{5.4}$$

where T_E and T_A are estimated durations of risk estimation and auxiliary processes, respectively. Solving Eq. 5.4. yields $T_E = 1,65$ seconds and $T_A = 0,22$ seconds. The average completion time of the model is recorded to be less than 10 seconds which can be considered as a fair value considering the fact that the proposed model is designed for offline operation without an explicit time constraint.

Memory Requirement

The proposed model is also analyzed in terms of its memory requirement by considering the memory usage during operation as well as size of its significant elements such as the patient dataset or output (result) files. The memory used by the algorithm is obtained by tracking the memory usage via Matlab's "memory" function during operation and subtracting the baseline memory usage of Matlab from the recorded values. The result of this analysis shows that the GUI uses a fixed 10 MB from the memory. The additional memory usage during runtime varies from 5 MB during display of model accuracy and 48 MB during IHTRE process. Therefore, the peak memory usage of the proposed model can be estimated as 58 MB which can be considered as next to negligible considering the computational power of modern machines. It should be noted that the method used to obtain these results provides an approximation of the peak memory usage of the model. The exact usage of memory depends on the memory allocation rules of Matlab, as well as the operating system itself, calculation of which is considerably tedious and outside of the scope of this study.

Lastly, the amount of storage required for significant elements of the proposed model is observed simply by recording the file size associated with each element. The result of this observation is given in Table 5.12.

Table 5.12. Sizes of significant elements of the proposed model

Element	File Size
Patient Dataset (.xlsx format)	2,87 MB
Patient Dataset (.mat format)	4,38 MB
Results File (.mat format)	≤ 5 MB
Clustering Algorithm	2 kB
Stroke Risk Estimation Algorithm	83 kB
GUI	61 kB
Total	≤ 9,53 MB

As can be seen from Table 5.12, a large portion of the occupied storage is not due to the model itself but its input and output files. The size of the results file varies between 4 and 5 MB depending on the sparsity of the result matrices contained within. For most cases, the size of this file is observed to be around 4,40 MB. The total storage required for

the proposed model is therefore calculated to be substantially small; less than 10 MB excluding the Patient Dataset in .xlsx format.

6. CONCLUSION

Stroke is a serious health condition that is considered among leading causes of disability and death worldwide. Despite this fact, the risk estimation studies in this field are observed to be relatively insufficient both by means of quality and quantity. This can be attributed to the acute nature of stroke and the long, incomplete list of risk factors; making it difficult to implement a reliable risk estimation system. This obstacle is aimed to be overcame by considering a comprehensive list of risk factors (features) and proposing a robust model that can accurately estimate the risks of four different types of stroke. By this way, the patients with a high risk of stroke can be determined accordingly.

In this study, an artificial intelligence based expert system that considers the risk factors of patients as 124 different input features and performs risk estimation regarding 4 different types of stroke is proposed. The proposed model consists of two modules in a cascade with the first module used to modify the input dataset with FCM clustering. In this module, the samples with known and unknown output features are used together in the clustering process which provides the model more resistance against inhomogeneously biased data. In the second module, four classifier ensembles consisting of four base models each are employed to perform risk estimation. The weights of base models of each ensemble are adjusted during model operation, according to the validation accuracy of the training dataset which is provided alongside the dataset of patients whose stroke risk is to be estimated. This dynamic approach in weight calculation provided a substantial decrease in the computational cost of utilizing a total of 16 classifiers while maintaining high estimation accuracy. Instead of settling for a readily available dataset, a new, comprehensive dataset of 3000 patients is formed from the raw data provided by Gazi University Hospital and used throughout the study. To enhance the comprehensibleness of the model outputs, as well as providing a user friendly interface, the proposed model is complemented with a fully functional GUI.

It is suggested that valuable contribution to the related literature is provided by a number of novel properties regarding the materials and methods associated with this study. These novel properties can be summarized in three main points as follows:

- An original and extensive dataset is formed using retrospectively obtained patient data and used within the scope of the study.
- A novel approach is followed in the two stage design of the proposed model which consists of a joint clustering stage followed by the risk estimation stage employing ensembles of ensembles alongside individual classifiers.
- A novel concept of dynamically adjusted weights that improve the model performance without increasing computational complexity is introduced and integrated to the model operation.

The performance of the proposed model is evaluated using the dataset formed in this study with a number of metrics taken under consideration with accuracy selected as the ultimate criteria. The accuracy of the proposed model associated with each stroke type is obtained as follows:

- General Stroke: 99,89%
- Hemorrhagic Stroke: 99,81%
- Ischemic Stroke: 99,90%
- TIA: 99,87%

While another study considering the risk estimations for four stroke types as given above was not found in literature, a number of studies involved estimation of stroke risk with their own definitions, the results of which can be compared to those obtained in the general stroke risk estimation part of this study.

When the results of the proposed model is compared to those of guideline based algorithms such as 2013 ACC/AHA, it is observed that the guideline based models were failed to provide accurate results, with the accuracy of 2013 ACC/AHA based model stated as 62,7%. The accuracy of the proposed model was obtained to be more than 35% higher than this value, not falling below 99,81% in estimation of risk associated with any stroke type. This is considered as an expected result as guideline based algorithms were shown to be inferior to those employing machine learning techniques as discussed in Section 2.3.

The proposed model clearly outperformed early studies in stroke risk estimation field such as Microstroke by Spitzer et. al. which was able to attain an overall accuracy value of 72,8%. However, it should be kept in mind that such early studies were carried out approximately three decades ago and therefore should be evaluated according to conditions of their period. Khosla et. al. proposed a model to predict the risk of stroke in 2010, the results of which can be compared to the results obtained for the base models used in general stroke risk estimation in this study as the performance of their model was expressed by AUC values. The best performing variant of the model proposed by Khosla et. al. was MCR with Conservative Mean used as the feature selection method, achieving 0,777 AUC. On the other hand, the base models used in the ensemble implemented for general stroke risk estimation were shown to perform much better, with AUC values between 0,91 and 0,94. Therefore, it can be claimed that the proposed model outperformed the model created by Khosla et. al by a large margin considering even the base models used in this study were able to significantly outperform their best performing variant.

Weng et. al. carried out a study to label patients with a high risk of CVD, including stroke in which the performance was expressed in terms of AUC values. Their best performing model, simply named as “Neural Network” was able to achieve an AUC of 0,764 which was lower than that achieved by the model proposed by Khosla et. al. When this value is compared to the AUC values obtained for base models used in the general stroke risk estimation ensemble of the proposed model, it is observed that the base models alone were able to clearly overperform the model proposed by Weng et. al. Considering this fact, the model proposed in this study has shown a significantly greater performance in estimating the risk of general stroke in patients while also providing the risk estimates for 4 types of stroke which were not analyzed in other studies in literature.

As can be seen from these comparisons, it was not possible to compare the proposed model to another model directly because of two main reasons. Firstly, the studies in literature did not consider the different types of stroke in the same way this study considered, which prevented a comparison of performance regarding stroke types. Secondly, the performance metric used in the studies in literature was mostly AUC, which was largely avoided in this study due to its unability to provide an objective assessment of model performance [190].

The performance of the system is further evaluated using F1 and F2 scores alongside confusion matrices obtained for each stroke type individually, which were provided in Section 5.2. As these results, alongside the comparisons provided above demonstrate, the proposed model achieved almost perfect accuracy for all four stroke types, setting a new benchmark in stroke risk estimation problem.

Another superiority of the model proposed in the study is its considerable resistance to bias in the dataset which may originate from external factors such as equipment based differences in measurements, thanks to the joint clustering employed in the first stage. Furthermore, the model is shown to be robust against variations in the dataset size and is able to preserve most of its high accuracy even with considerably small training datasets.

The model's performance was also evaluated in terms of memory and time as discussed in Section 5.2.2. Its operation can be considered to be fairly quick with risk estimation of up to 3000 patients completed in less than 10 seconds on average. Furthermore, the runtime memory requirement for the model operation is calculated to be substantially small, not exceeding 58 MB during peak operation. The storage space required in memory for the files necessary for model operation was also observed to be less than 10 MB which should be of little concern considering the storage space available in modern computers. It should be however, noted that the model does not run as a standalone executable file but as an application in Matlab, which takes up several gigabytes of space in memory.

It should be kept in mind that as discussed in Section 2.1.4, there are a number of risk factors that affect the risk of different types of stroke differently. Therefore, the precautions that should be taken may vary depending on the stroke type as well as the evaluation of the patient regarding the risk of different stroke types. By informing the user about the type of stroke event that is more likely to occur alongside providing the risk estimations for each stroke type, the proposed model makes it easier to decide on the preventive measures to be taken for each patient individually. Considering that around 80% of stroke events are preventable by taking the necessary precautions as suggested by World Health Association [191], it is believed that the model proposed with this study can be a useful tool for prevention of such events. In this context, implementing a robust model that can successfully aid the medical experts in saving precious lives by acting as a decision support system is regarded as the ultimate goal of this study. Therefore, it is believed that an interdisciplinary

study such as this thesis will provide valuable contribution to both medical and engineering literature with its methods and results.

Stroke risk estimation is considered as a research area of utmost importance considering the number of lives affected by stroke attacks worldwide. It is, therefore, not possible nor desirable to limit the solutions proposed in this field to a single study. However, a number of recommendations for future studies can be made considering the present state of the model proposed in this study as provided below:

- While being from one of the largest and most advanced university hospitals in Turkey, the dataset used in the study is still a single-center one. Single center studies are generally considered to be prone to contradicting with similar studies involving other data sources. Therefore, forming a more diverse dataset by obtaining patient data from multiple centers would be preferable in the absence of certain limitations such as time or bureaucracy.
- Currently, the patient data must be manually formed with a fixed order of features either as a MS Excel (.xls or .xlsx) or Matlab (.mat) file. An automatic data acquisition feature can be implemented to build the patient dataset from raw data to speed up the process.
- The proposed model is implemented as a Matlab application instead of a standalone (MS Windows) executable. Therefore, in order to use the application, Matlab must be installed on a personal computer and a very basic level knowledge of Matlab might be necessary to launch the model's application. For improved convenience, the model can be implemented as a standalone application for computers as well as mobile phones. The mobile application can be implemented as the client side of a client-server network in which the client uploads the patient data and receives the results from the server which carries out the risk estimation task.
- Using the dataset formed in the study, the association of each risk factor (feature) with the risk of each stroke type can be investigated in detail as a retrospective cohort study. By this way, it can be possible to discover new risk factors with previously unknown effects on stroke risk.

REFERENCES

1. Adams, H. P., Adams, R. J., Brott, T., Zoppo, G. J. D., Furlan, A., Goldstein, L. B., Grubb, R. L., Higashida, R., Kidwell, C., Kwiatkowski, T. G., Marler, J. R., and Hademenos, G. J. (2003). Guidelines for the early management of patients with ischemic stroke: a scientific statement from the Stroke Council of the American Stroke Association. *Stroke,* 34(4), 1056-1083.

2. Li, S., Nie, E. H., Yin, Y., Benowitz, L. I., Tung, S., and Vinters, H. V. (2015). GDF10 is a signal for axonal sprouting and functional recovery after stroke. *Nature Neuroscience,* 18, 1737-1745.

3. Hicks, K. A., Mahaffey, K. W., Mehran, R., Nissen, S. E., Wiviott, S. D., Dunn, B., Solomon, S. D., Marler, J. R., and Teerlink, J. R. (2018). 2017 Cardiovascular and stroke endpoint definitions for clinical trials. *Circulation,* 137(9), 961-972.

4. Makris, K., Haliassos, A., Chondrogianni, M., and Tsivgoulis, G. (2018). Blood biomarkers in ischemic stroke: potential role and challenges in clinical practice and research. *Critical Reviews in Clinical Laboratory Sciences,* 55(5), 294-328.

5. Wang, Q., Reps, J. M., Kostka, K. F., Ryan, P. B., Zou, Y., and Voss, E. A. (2020). Development and validation of a prognostic model predicting symptomatic hemorrhagic transformation in acute ischemic stroke at scale in the OHDSI network. *Plos One,* 15(1).

6. Wadley, V. G., McClure, L. A., Howard, V. J., Unverzagt, F. W., Go, R. C., and Moy, C. S. (2007). Cognitive status, stroke symptom reports, and modifiable risk factors among individuals with no diagnosis of stroke or transient ischemic attack in the reasons for geographic and racial differences in stroke (REGARDS) study. *Stroke,* 38(4), 1143-1147.

7. Weng, S., F., Reps, J., Kai, J., Garibaldi, J., M., and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *Plos One,* 12(4).

8. Paciaroni, M., Bandini, F., Agnelli, G., Tsivgoulis, G., Yaghi, S., and Furie, K. L. (2018). Hemorrhagic transformation in patients with acute ischemic stroke and atrial fibrillation: time to initiation of oral anticoagulant therapy and outcomes. *Journal of the American Heart Association,* 7(22), 4-5.

9. Tait, R. J, Caldicott., D., Mountain, D., Hill, S. L., and Lenton, S. (2015). A systematic review of adverse events arising from the use of synthetic cannabinoids and their associated treatment. *Clinical Toxicology,* 54(1), 1-13.

10. World Stroke Organization. (2017). *Global stroke fact sheet.* Geneva: World Stroke Organization, 3-11.

11. Krishnamurthi, R. V., Feigin, V. L., Forouzanfar, M. H., Mensah, G. A., Connor, M., and Bennett, D. A. (2013). Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990-2010: findings from the global burden of disease study 2010. *The Lancet Global health,* 1(5), 259-281.

12. Internet: Stroke Facts and Statistics. (2020). URL: https://www.strokeinfo.org/stroke-facts-statistics/, Last Accessed: 13 May 2021.

13. Fisher, M., Moores, L., Alsharif, M. N., and Paganini, A. (2016). Definition and implications of the preventable stroke. *Journal of American Medical Association Neurology,* 73(2), 186-189.

14. Orfanoudaki, A., Chesley, E., Cadisch, C., Stein, B., Nouh, A., Alberts, M., and Bertsimas, D. (2020). Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score. *Plos One,* 15(5).

15. Raichle, M. E. (1983). The pathophysiology of brain ischemia. *Annals of Neurology*, 13(1), 2-10.

16. Kasper, D. L., Braunwald, E., Fauci, A. S., Hauser, S. L., Longo, D. L., and Jameson J. L. (Editors). (2005). *Harrison's principles of internal medicine* (16th Edition). New York: McGraw-Hill Medical, 2559-2570.

17. World Health Organisation. (1978). *Cerebrovascular disorders.* Geneva: World Health Organisation Offset Publications.

18. Internet: Types of Stroke. (2021). URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/types-of-stroke, Last Accessed: 13 May 2021.

19. Shuaib, A., and Hachinski, V. C. (1991). Mechanisms and management of stroke in the elderly. *Canadian Medical Association Journal,* 145(5), 433-443.

20. Donnan, G. A., Fisher, M., Macleod, M., and Davis, S. M. (2008). Stroke. *The Lancet*, 371(9624), 1612-1623.

21. Adams, H. P., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., and Marsh, E. E. (1993). Classification of subtype of acute ischemic stroke, definitions for use in a multicenter clinical trial TOAST, Trial of Org 10172 in Acute Stroke Treatment. *Stroke,* 24(1), 35-41.

22. Courbier, R. (1985). *Basis for a classification of cerebral arterial diseases.* Amsterdam: Excerpta Medica, 308.

23. Hacke, W., Zoppo, D. G. J., and Harker, L. A. (1987). Thrombosis and cerebro-vascular disease. In K. Poeck, E. B. Ringelstein and W. Hacke (Eds.), *Recent advances in diagnosis and management of stroke.* Berlin: Springer, 59.

24. Levy, D. E. (1988). How transient are transient ischemic attacks? *Neurology*, 38(674), 14-19.

25. Kleindorfer, D., Panagos, P., Pancioli, A. Khoury, J., Kissela, B., Woo, D., Schneider, A., Moomaw, C., Shukla, R., and Broderick, J. P. (2005). Incidence and short-term prognosis of transient ischemic attack in a population-based study. *Stroke*, 36(4), 720-723.

26. Liliang, P. C., Liang, C. L., and Lu, C. H. (2001). Hypertensive caudate hemorrhage prognostic predictor, outcome, and role of external ventricular drainage. *Stroke*, 32(5), 1195-1200.

27. Klein, J., Pohl, J., Vinson, E. N., Brant, W. E., and Helms, C.A. (2018). *Brant and Helms' fundamentals of diagnostic radiology*. (5th Edition). Alphen aan den Rijn, Netherlands: Wolters Kluwer, 195.

28. Broderick, J. P., Adams, H. P., Barsan, W., Feinberg, W., Feldmann, E., Grotta, J., Mayberg, M., Tilley, B., Zabramski, J. M., and Zuccarello, M. (1999). Guidelines for the management of spontaneous intracerebral hemorrhage: A statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. *Stroke*, 30(4), 905-915.

29. Ögren, J., Irewall, A. L., Bergström, L., and Mooe, T. (2015). Intracranial hemorrhage after ischemic stroke incidence, time trends, and predictors in a Swedish nationwide cohort of 196 765 patients. *Circulation: Cardiovascular Quality and Outcomes*, 8(4), 413-420.

30. Trouillas, P., and Kummer, R. V. (2006). Classification and pathogenesis of cerebral hemorrhages after thrombolysis in ischemic stroke. *Stroke*, 37(2), 556-561.

31. Molina, C., Alvarez-Sabin, J., Montaner, J., Abilleira, S., Arenillas, J. F., and Coscojuela, P. (2002). Thrombolysis-related hemorrhagic infarction. A marker of early reperfusion, reduced infarct size, and improved outcome in patients with proximal middle cerebral artery occlusion. *Stroke*, 33(6), 1551-1556.

32. Feigin, V. L., Nguyen, G., Cercy, K., Johnson, K. O., Alam, T., Parmar, P. G., Abajobir, A. A., Abate, K. H., Allah, F., Abejie, A. N., Abyu, G.Y., and Akinyemi, R. O. (2018). Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016. *New England Journal of Medicine*, 379(25), 2429-2437.

33. Snyder, B. D., and Lassepas, R. M. (1980). Cerebral infarction in young adults. *Stroke*, 11, 149-153.

34. Leno, C., Berciano, J., Combarros, O., Polo, J. M., Pascual, J., Quintana, F., Merino, J., Duran, R. M., and Alvarez, C. (1993). A prospective study of stroke in young adults in Cantabria, Spain. *Stroke*, 24, 792-795.

35. Federico, F., Calvario, T., Turi, D. N., and Paradiso, F. (1990). Ischaemic cerebral infarction in young adults. *Acta Neurology*, 12, 101-108.

36. Castellano, M. (2009). Conventional risk factors. In A. Pezzini and A. Padovani (Eds.), *Cerebral ischemia in young adults.* New York: Nova Science, 25.

37. Franklin, S. S., Wong, N. D., and Kannel, W. B. (2003). Age-specific relevance of usual blood pressure to vascular mortality. *The Lancet*, 361(9366), 1389.

38. Sarwar, N., Gao, P., Seshasai, S. R., Gobin, R., Kaptoge, S., Angelantonio, D., Ingelsson, E., Lawlor, D. A., Selvin, E., and Stampfer, M. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Emerging risk factors collaboration. *The Lancet*, 375(9733), 2215-2222.

39. Internet: Metzger, B. (2021). *Gestational Diabetes*. National Institute of Diabetes and Digestive and Kidney Diseases. URL: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/gestational, Last Accessed: 14 May 2021.

40. Falconer, J. A., Naughton, B. J., Strasser, D. C., and Sinacore, J. M. (1994). Stroke inpatient rehabilitation: a comparison across age groups. *Jam Geriatry,* 42, 39-44.

41. Pohjasvaara, T., Erkinjuntti, T., Vataja, R., and Kaste, M. (1997). Comparison of stroke features and disability in daily life in patients with ischemic stroke aged 55 to 70 and 71 to 85 years. *Stroke,* 28, 729-735

42. Jorgensen, H., Nakayama, H., Raaschou, H. O., and Olsen, T. S. (1994). Stroke in patients with diabetes: the Copenhagen Stroke Study. *Stroke,* 25, 1977-1984.

43. Lehto, S., Ronnemaa, T., Pyorala, K., and Laakso, M. (1996). Predictors of stroke in middle-aged patients with non-insulin-dependent diabetes. *Stroke,* 27, 63-68.

44. Davis, T. M., Millns, H., Stratton, I. M., Holman, R. R., and Turner, R. C. (1999). Risk factors for stroke in type 2 diabetes mellitus: United Kingdom prospective diabetes Study (UKPDS). *Archieves of Internal Medicine,* 159, 1097-1103.

45. Folsom, A. R., Rasmussen, M. L., Chambless, L. E., Howard, G., Cooper, L. S., Schmidt, M. I., and Heiss, G. (1999). Prospective associations of fasting insulin, body fat distribution, and diabetes with risk of ischemic stroke: the Atherosclerosis Risk in Communities (ARIC) Study investigators. *Diabetes Care,* 22, 1077-1083.

46. Janghorbani M., Hu, F. B., and Willett, W. C. (2007). Prospective study of type 1 and type 2 diabetes and risk of stroke subtypes: the nurses' health study. *Diabetes Care,* 30, 1730.

47. Stahl, C. H., Lind, M., Svensson, A. M., Gudbjörnsdottir, S., Martensson, A., and Rosengren, A. (2017). Glycaemic control and excess risk of ischaemic and haemorrhagic stroke in patients with type 1 diabetes: a cohort study of 33 453 patients. *Journal of Internal Medicine,* 281(3), 261-272.

48. Grysiewicz, R., Thomas, K., and Pandey, D. K. (2008). Epidemiology of ischemic and hemorrhagic stroke: incidence, prevalence, mortality, and risk factors. *Neurologic Clinics,* 26(4), 871-95.

49. Agostino, D. R. B., Wolf, P. A., Belanger, A. J., and Kannel, W. B. (1994) Stroke risk profile: adjustment for antihypertensive medication: the Framingham Study. *Stroke,* 25, 40-43.

50. Internet: Difference Between Dyslipidemia and Hyperlipidemia. (2017, Aug). URL: https://www.differencebetween.com/difference-between-dyslipidemia-and-vs-hyperlipidemia. Last Accessed: 14 May 2021.

51. Kannel, W. B., Dawber, T. R., Cohen, M. E., and McNamara P. M. (1965). Vascular disease of the brain-epidemiologic aspects: the Framingham Study. *Maj Public Health,* 55, 1355-1366.

52. Iso, H., Jacobs D. R., Wentworth, D., Neaton, J. D., and Cohen, J. D. (1989). Serum cholesterol levels and six-year mortality from stroke in 350,977 men screened for the multiple risk factor intervention trial. *New England Medicine,* 320, 904-910.

53. Bots, M. L., Elwood, P. C., Nikitin, Y., Salonen, J. T., Concalves F. D. A., Intizari D., Trichopoulou, A., Tuomilehto, J., Koudstaal, P. J., and Grobbee, D. E. (2002). Total and HDL cholesterol and risk of stroke. EUROSTROKE: a collaborative study among research centres in Europe. *Epidemiology Community Health,* 56(1), 19-24.

54. Zhang, X., Patel, A., Horibe, H., Wu, Z., Barzi, F., Rodgers, A., and Macmahon, S. (2003). Cholesterol, coronary heart disease, and stroke in the Asia Pasific region. *International Journal of Epidemiology,* 32, 563-572.

55. Tirschwell, D. L., Smith, N. L., Heckbert, S. R., Lemaitre, R. N., Longstreth, W. T., and Psaty, B. M. (2004). Association of cholesterol with stroke risk varies in stroke subtypes and patient subgroups. *Neurology,* 63, 1868-1875.

56. Kagan, A., Popper, J. S., and Rhoads, G. G. (1980). Factors related to stroke incidence in Hawaiian Japanese men, the Honolulu heart study. *Stroke,* 11, 14-21.

57. Prospective Studies Collaboration. (1995). Cholesterol, diastolic blood pressure, and stroke: 13 000 strokes in 450 000 people in 45 prospective cohorts. *The Lancet,* 346, 1647-1653.

58. Yano, K., Reed, D. M., and Maclean, C. J. (1989). Serum cholesterol and hemorrhagic stroke in the Honolulu heart program. *Stroke,* 20(11), 1460-1465.

59. Eastern Stroke and Coronary Heart Disease Collaborative Research Group. (1998). Blood pressure, cholesterol, and stroke in eastern Asia. *The Lancet,* 352, 1801-1807.

60. Dökmeci, İ., and Dökmeci, A. H. (2011). *Tıp terimleri sözlüğü* (2nd ed.). İstanbul: İstanbul Medikal Yayıncılık.

61. Amarenco, P., Goldstein, L. B., Messig, M., O'Neill, B. J., Callahan, A., Sillesen, H., Simunovic, L., Zivin, J. A., and Welch, K. M. A. (2009). Relative and cumulative effects of lipid and blood pressure control in the stroke prevention by aggressive reduction in cholesterol levels trial. *Stroke,* 40(7), 2486-2492.

62. Griffin, M. J., and Hines, R. L. (2001). Management of perioperative ventricular dysfunction. *Cardiothoracic Vasculo-anesthesia,* 15(1), 90-106.

63. Szummer, K. E., Solomon, S. D., Velazquez, E. J., Kilaru, R., McMurray, J., and Rouleau, J. L. (2005). Heart failure on admission and the risk of stroke following acute myocardial infarction: the VALIANT registry. *Euro Heart,* 26(20), 2114-2119.

64. Sundboll, J., Puho, E. H., Schmidt, M., Pedersen, L., Henderson V. W., Botker, H. E., and Sorensen, H. T. (2016). Long-term risk of stroke in myocardial infarction survivors thirty-year population-based cohort study. *Stroke,* 47, 1727-1733.

65. Goldstein, L. B., Amarenco, P., Szarek, M., Callahan, A., Hennerici, M., Sillesen, H., and Welch, K. M. A. (2008). Hemorrhagic stroke in the stroke prevention by aggressive reduction in cholesterol levels study. *Neurology,* 70(24), 2364-2370.

66. Habib, G. (2003). Embolic risk in subacute bacterial endocarditis: determinants and role of transesophageal echocardiography. *Cure Cardiology,* 5(2), 129-136.

67. Yellapu, V., Ackerman, D., Longo, S., and Stawicki, S. P. (2018). Septic embolism in endocarditis: anatomic and pathophysiologic considerations. In M. S. Firstenberg (Editor), *Advanced concepts in endocarditi*s. London: IntechOpen, 150-170.

68. Katsouli, A., and Massad, M. G. (2013). Current issues in the diagnosis and management of blood culture negative infective and non-infective endocarditis. *The Annals of Thoracic Surgery,* 95(4), 1467-1474.

69. Steckelberg, J. M., Murphy, J. G., Ballard, D., Bailey, K., Tajik, A. J., Taliercio, C. P., and Wilson, W. R. (1991). Emboli in infective endocarditis: the prognostic value of echocardiography. *Ann Intern Medicine,* 114(8), 635-640.

70. Ruttman, E., Willeit, J., Ulmer, H., Chevtchik, O., Höfer, D., Poewe, W., Laufer, G., and Müller, L. C. (2006). Neurological outcome of septic cardioembolic stroke after infective endocarditis. *Stroke,* 37(8), 2094-2099.

71. Derex, L., Bonnefoy, E., and Delahaye, F. (2010). Impact of stroke on therapeutic decision making in infective endocarditis. *Journal of Neurology,* 57(3), 315-321.

72. Katz, E., Katz, S. B., Fedorchuk, C., Lightstone, D. F., Banach, C. J., and Podoll, J. (2019). Increase in cerebral blood flow indicated by increased cerebral arterial area and pixel intensity on brain magnetic resonance angiogram following correction of cervical lordosis. *Brain Circulation,* 5(1), 19-26.

73. Heiro, M., Nikoskelainen, J., Engblom, E., Kotilainen, E., Marttila, R., and Kotilainen, P. (2000). Neurologic manifestations of infective endocarditis a 17-year experience in a teaching hospital in Finland. *Archieves of Internal Medicine,* 160(18), 2781-2787.

74. Karp, R. B., Cyrus, R. J., Blackstone, E. H., Kirklin, J. W., Kouchoukos, N.T., and Pacifico, A. D. (1981). The Björk-Shiley valve: intermediate-term follow-up. *Journal of Thoracic Cardiovascular Surgery*, 81(4), 602–614.

75. Daenen, W., Nevelsteen, A., Cauwelaert, P., Maesschalk, E., Willems, J., and Stalpaert, G. (1983). Nine years' experience with the Björk-Shiley prosthetic valve: early and late results of 932 valve replacements. *Ann Thoracic Surgery*, 35(6), 651–663.

76. Lepley, D., Flemma, R. J., Mullen, D. C., Singh, H., and Chakravarty, S. (1977). Late evaluation of patients undergoing valve replacement with the Björk-Shiley prosthesis. *Ann Thoracic Surgery*, 24(2), 131–139.

77. Cabral, M. R. J., McNamara, J. J., Mamiya, R. T., Brainard, S. C., and Chung, G. K. T. (1978). Acute thrombotic obstruction with Björk-Shiley valves: diagnostic and surgical considerations. *Thoracic Cardiovascular Surgery*, 75, 321-329.

78. Roy, D., Marchand, E., Gagne, P., Chabot, M., and Cartier, R. (1986). Usefulness of anticoagulant therapy in the prevention of embolic complications of atrial fibrillation. *Heart Journal*, 112(5), 1039-1043.

79. Vahanian, A., Andreotti, F., Antunes, M. J., Esquivias, G. B., Baumgartner, H., Borger, M. A., Carrel, T. P., and Bonis, M. D. (2012). Guidelines on the management of valvular heart disease (version 2012): The joint task force on the management of valvular heart disease of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). *European Heart Journal*, 33(19), 2451-2496.

80. Supple, G. E., Ren, J. F., Zado, E. S., and Marchlinski, F. E. (2011). Mobile thrombus on device leads in patients undergoing ablation. *Circulation*, 124, 772-778.

81. U.S. Department of Health and Human Services. (2020). *Smoking cessation: a report of the surgeon general.* Rockville, U.S.A.: U.S. Department of Health and Human Services, 92-127.

82. U.S. Department of Health, Education, and Welfare. (1964). *Smoking and health: report of the advisory committee to the surgeon general of the public health service*. Washington: U.S. Department of Health, Education, and Welfare, Public Health Service.

83. Shinton, R., and Beevers, G. (1989). Meta analysis of relation between cigarette smoking and stroke. *British Medical Journal*, 298, 789-794.

84. Lipton, R. B., and Silberstein, S. D. (1994). Why study the comorbidity of migraine? *Neurology*, 44(7), 4-5.

85. Etminan, M., Takkouche, B., Isorna, F. C., and Samii, A. (2005). Risk of ischemic stroke in people with migraine: systematic review and meta-analysis of observational studies. *British Medical Journal*, 330, 63-65.

86. Regan, O. C., Wu, P., Arora, P., Perri, D., and Mills, E. J. (2008). Statin therapy in stroke prevention: a meta-analysis involving 121,000 patients. *American Journal of Medicine*, 121, 24-33.

87. Amarenco, P., Bogousslavsky, J., Callahan, A., Goldstein L. B., Hennerici, M., Rudolph, A. E., Sillesen, H., Simunovic, L., and Szarek, M. (2006). High-dose atorvastatin after stroke or transient ischemic attack. *New England Medicine*, 355(6), 549-559.

88. Chan, W. S., Ray, J., Wai, E. K., Ginsburg, S., Hannah, M. E., Corey, P. N., and Ginsberg, J. S. (2004). Risk of stroke in women exposed to low-dose oral contraceptives: a critical evaluation of the evidence. *Archives of Internal Medicine*, 164, 741-747.

89. Rinde, L. B., Smabrekke, B., Mathiesen, E. B., Lochen, M. L., Njolstad, I., and Hald, E. M. (2016). Ischemic stroke and risk of venous thromboembolism in the general population: the Tromsø Study. *Journal of the American Heart Association*, 5(11), 1-8.

90. Langhorne, P., Stott, D. J., Robertson, L., MacDonald, J., Jones, L., McAlpine, C., and Dick, F. (2000). Medical complications after stroke: a multicenter study. *Stroke,* 31, 1223–1229.

91. Internet: Anticoagulant Medications. (2021). URL: http://www.womens-health-advice.com/heart-disease/anticoagulants.html. Last Accessed: 14 May 2021.

92. Edwards, S. H. (2014). *Nonsteroidal anti-inflammatory drugs*. Kenilworth, U.S.A: Merck.

93. Sloan, M. (1997). Toxicity substance abuse. In K. Welch, L. Caplan and D. Reis (Eds.), *Primer on cerebrovascular diseases.* San Diego: Academic Press, 413-416.

94. Levine, S. R., Brust, J. C., and Futrell, N. (1991). A comparative study of the cerebrovascular complications of cocaine; alkaloidal versus hydrochloride. *Neurology,* 141, 1173-1177.

95. Gericke, O. L. (1945). Suicide by ingestion of amphetamine sulfate. *Journal of American Medical Applications,* 28, 1098-1099.

96. Westover, A. N., McBride, S., and Haley, R. W. (2007). Stroke in young adults who abuse amphetamines or cocaine: a population-based study of hospitalised patients. *Arch Psychiatry,* 64, 495-502.

97. Brust, J. C., and Richter, R. W. (1976). Stroke associated with addiction to heroin. *Neurology Neurosurgery Psychiatry*, 39, 194-199.

98. Woods, B. T., and Strewler, G. J. (1972). Hemiparesis occuring six hours after intravenous heroin injection. *Neurology,* 22, 863-866.

99. Herskowitz, A., and Gross, E. (1973). Cerebral infarction associated with heroin sniffing. *South Medical Journal*, 66, 778-784.

100. Vila, N., and Chamorro, A. (1997). Ballistic movements due to ischaemic infarcts after intravenous heroin overdose: report of two cases. *Neurology Neurosurgery,* 99, 259-262.

101. Fisher, B. A. C., Ghuran, A., and Vadamalai, V. (2005). Cardiovascular complications induced by cannabis smoking: a case report and review of the literature. *Emergency Medical Journal,* 22, 679-680.

102. Singh, G. K. (2000). Atrial fibrillation associated with marijuana use. *Pediatric Cardiology,* 21, 284.

103. Sobel, J., Espinas, O. E., and Friedman, S. A. (1971). Carotid artery obstruction following LSD capsule ingestion. *Arch Intern Medicine,* 127, 290-291.

104. Lieberman, A. N., Bloom, W., Kishore, P. S., and Lin, J. P. (1974) Carotid artery occlusion following ingestion of LSD. *Stroke,* 5, 213-215.

105. Rumbaugh, C. L., Bergeron, R. T., Fang, H. C., and McCormick, R. (1971) Cerebral angiographic changes in the drug abuse patient. *Radiology,* 101, 335-344.

106. Malarcher, A. M., Giles, W. H., Croft, J. B., Wozniak, M. A.,Wityk, R. J., Stolley, P. D., Sloan, M. A., Sherwin, R., Price, T. R., Macko, R. F., and Johnson, C. J. (2001). Alcohol intake, type of beverage, and the risk of cerebral infarction in young women. *Stroke,* 32(1), 77-83.

107. Gaziano, J. M., Buring, J. E., Breslow, J. L., Goldhaber, S. Z., Rosner, B., Vandenburgh, M., and Hennekens, C. H. (1993). Moderate alcohol intake, increased levels of high-density lipoprotein and its subfractions, and decreased risk of myocardial infarction. *New England Journal of Medicine,* 329, 1829-1834.

108. Langer, R. D., Criqui, M: H., and Reed, D. M. (1992). Lipoproteins and blood pressure as biological pathways for effect of moderate alcohol consumption on coronary heart disease. *Circulation,* 85, 910-915.

109. Suh, I., Shaten, B. J., Cutler, J. A., and Kuller, L. H. (1994). Alcohol use and mortality from coronary heart disease: the role of high-density lipoprotein cholesterol. the multiple risk factor intervention trial research group. *Annual Intern Medicine,* 116, 881-887.

110. Feldmann, E., Broderick, J. P., Kernan, W. N., Viscoli, C. M., Brass, L. M., Brott, T., Wilterdink, J. L., and Horwitz, R. I. (2005). Major risk factors for intracerebral hemorrhage in the young are modifiable. *Stroke,* 36(9), 1881-1885.

111. Hardalaç, F., Kutbay, U., and Salimitorkamani, M. (2016). *Patient diagnosis with ECG signals estimation using artificial neural network.* International Conference on Engineering and Technology, Computer, Basic and Applied Sciences. Dubai, 13.

112. Livingstone, D. J. (2008). *Artificial neural networks methods and applications.* Sandown, UK: Springer, v-6.

113. McCorduck, P. (2004). *Machines who think: a personal inquiry into the history and prospects of artificial intelligence* (Second edition). Natick, U.S.A: A K Peters, Ltd., 335, 434–435.

114. Newell, A., and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the Association for Computing Machinery,* 19(3), 113-126.

115. MacLennan, B. (2009). History of artificial intelligence before computers. In *Encyclopedia of information science and technology.* Second Edition. Hershey: IGI Global, 1763–1768.

116. Maio, P. D. (2019). Neurosymbolic knowledge representation for explainable and trustworthy AI. *Systems,* 7, 2-22.

117. Burkert, W. (1972). *Lore and science in ancient pythagoreanism.* Nürnberg: Hans Carl, 496.

118. Burnet, J. (1920). *Early Greek philosophy.* London: A & C Black Ltd., 301-320

119. Bartheld, C. S. V., Bahney, J., and Houzel, S. H. (2016). The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Computer Neurology,* 524(18), 3865-3895.

120. Purves, D. (2008). *Neuroscience.* (Fourth edition). Sunderland, U.S.A: Sinauer Associates, 526.

121. Rizo, J., and Rosenmund, C. (2008). Synaptic vesicle fusion. *Nature Structural & Molecular Biology*, 15, 665-674.

122. Russell, S. J., and Norvig, P. (2003). *Artificial intelligence: a modern approach.* (Second edition). New Jersey: Prentice Hall, 189-201, 216-222.

123. Serhatlıoğlu, S., Hardalaç, F., and Güler, İ. (2003). Classification of transcranial Doppler signals using artificial neural network. *Journal of Medical Systems,* 27(2), 205-214.

124. Mitchell, T. M. (1997). *Machine learning*. (First edition). New York: McGraw-Hill Education, 162-164.

125. Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210-229.

126. Nilsson, N. J. (1965). *Learning machines: foundations of trainable pattern-classifying systems.* New York: McGraw-Hill, 16-28.

127. Langey, P. (2011). The changing science of machine learning. *Machine Learning*, 82(3), 275-279.

128. Lanka, P., Rangaprakash, D.., Dretsch, M. N., Katz, J. S., Denney, T. S., and Deshpande, G. (2020). Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. *Brain Imaging and Behavior,* 14, 2378-2416.

129. Apaydın, E. (2010). *Introduction to machine learning.* (Second edition). Cambridge: MIT Press, 9, 155, 202.

130. Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100-108.

131. İnkaya, T. (2015). A density and connectivity based decision rule for pattern classification. *Expert Systems with Applications,* 42(2), 906-912.

132. Dang, X., Assent, I., and Bailey, J. (2012). Multiple clustering views via constrained projections. *3rd Multiclass Workshop,* 23-30.

133. Hajiramezanali, E., Dadaneh, S. Z., Karbalayghareh, A., Zhou, M., and Qian, X. (2018). *Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data*. 32nd Conference on Neural Information Processing Systems. Montreal, Canada, 1-10.

134. Hooper, D. F. (1988). *SID: synthesis of integral design*. Proceedings of the 1988 IEEE International Conference on Computer Design: VLSI. New York, 204-208.

135. Gibson, C. S. (1990). VAX 9000 series. *Digital Technical Journal of Digital Equipment Corporation,* 2(4), 118-129.

136. Roth, F. H., and Lesser, V. R. (1977). Focus of attention in the Hearsay-II system. *Natural Language,* 2, 27-35.

137. Roth, F. H., Waterman D. A., and Lenat, D. B. (1983). *Building expert systems.* Reading: Addison-Wesley Pub. Co., 309-310.

138. Pułka A., and Kłosowski P. (2009). Polish speech processing expert system incorporated into the EDA tool. In: Z. S. Hippe and J. L. Kulikowski (Eds.) *Human computer systems interaction.* Berlin: Springer, 60-63.

139. Woolery, L. K., and Busse, G. J. (1994). Machine learning for an expert system to predict preterm birth risk. *Journal of the American Medical Informatics Association,* 1(6), 439-446.

140. Lee, K. H., and Jo, G. S. (1999). Expert system for predicting stock market timing using a candlestick chart. *Expert Systems with Applications,* 16(4), 357-364.

141. Kwak, S. H., Ong, S. M., and McGhee, R. B. (1990). A mission planning expert system for an autonomous underwater vehicle. *Symposium on Autonomous Underwater Vehicle Technology*, 123-128.

142. Sunberg, Z., and Rogers, J. (2013). A fuzzy logic based controller for helicopter autorotation. *51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition,* 1-23.

143. Sunberg, Z., Miller, N. R., and Rogers, J. (2015). A real time expert control system for helicopter autorotation. *Journal of the American Helicopter Society,* 60(2), 1-15.

144. Internet: UCLA Stroke Center. (2019). Stroke Risk Calculator. URL: Https://Www.Uclahealth.Org/Stroke/Stroke-Risk-Calculator. Last Accessed: 27 Apr 2021.

145. Internet: Cleveland Clinic Neurosurgical Institute. (2019). Stroke Risk Calculator. URL: https://Clevelandclinic.Org/Departments/Neurological/Depts/Cerebrovascular/Stroke-Risk-Calculator. Last Accessed: 29 Apr 2021.

146. McGeechan, K., Liew, G., Macaskill, P., Irwig, L., Klein, R., Klein, B. E., and Sharrett, A. R. (2009). Prediction of incident stroke events based on retinal vessel caliber: a systematic review and individual-participant meta-analysis. *American Journal of Epidemiology,* 170(11), 1323-1332.

147. GoffJr, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., Lackland, D. T., Levy, D., Donnell, C. J., Robinson, J. G., and Smith, S. C. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation*, 129(25), 49-73.

148. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., and Sundberg, P. (2018). Scalable and accurate deep learning with electronic health records. *New Jersey Digital Medicine*, 1(18), 1-10.

149. Johnson, K. M., Johnson, H. E., Zhao, Y., Dowe, D. A., and Staib, L. H. (2019). Scoring of coronary artery disease characteristics on coronary CT angiograms by using machine learning. *Radiology*, 292(2), 354-362.

150. Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., and Protsyuk, I. V. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572, 116-119.

151. Spitzer, K., Thie, A., Caplan, L. R., and Kunze, K. (1989). The MICROSTROKE expert system for stroke type diagnosis. *Stroke*, 20(10), 1353-1356.

152. Siegel, J. S., Ramsey, L. E., Snyder, A. Z., Metcalf, N. V., Chacko, R. V., and Weinberger, K. (2015). Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. *Proceedings of the National Academy of Sciences of the United States of America Neuroscience*, 113(30), 4367-4376.

153. Park, E., Chang, H., and Nam, H. S. (2018). A Bayesian network model for predicting post-stroke outcomes with available risk factors. *Frontiers in Neurology*, 7(9), 699.

154. Khosla, A., Cao, Y., Lin, C. C., Chiu, H. K., Hu, J., and Lee, H. (2010*). An integrated machine learning approach to stroke prediction.* Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining. New York, 183-192.

155. Kuncheva, L. and Whitaker, C. (2003). Measures of diversity in classifier ensembles. *Machine Learning*, 51, 181-207.

156. International Stroke Trial Collaborative Group. (1997). The international stroke trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *The Lancet*, 349(9065), 1569-1581.

157. Asplund, K., Asberg, H. K., Norrving, B., Stegmayr, B., and Terent, A. (2003). Riksstroke, a Swedish national quality register for stroke care. *Cerebrovascular Diseases*, 15(1), 5-7.

158. Song, Y. M., Sung, J., Smith, G. D., and Ebrahim, S. (2004). Body mass index and ischemic and hemorrhagic stroke. *Stroke*, 35, 831-836.

159. Kernan, W. N., Inzucchi, S. E., Sawan, C., Macko, R. F., and Furie, K. L. (2012). Obesity a stubbornly obvious target for stroke prevention. *Stroke*, 44(1), 278-286.

160. Internet: International Alliance for Responsible Drinking (IARD). (2019). Drinking guidelines: general population. URL: https://iard.org/science-resources/detail/Drinking-Guidelines-General-Population, Last Accessed: 16 May 2021.

161. Hays, William L. (1994). *Statistics* (5th edition). Orlando, Florida: Harcourt Brace, 495.

162. Al-Razgan, M., and Domeniconi, C. (2006). *Weighted clustering ensembles*. Proceedings of the 2006 Siam International Conference on Data Mining. Philadelphia, U.S.A., 258-269.

163. Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2), 123-140.

164. Balogh, E. P., Miller, B. T., Ball, J. R., Committee on Diagnostic Error in Health Care; Board on Health Care Services; Institute of Medicine (2015). *Improving diagnosis in health care*. Washington: National Academies Press, 2.

165. Klawonn, F., Kruse, R., and Winkler, R. (2015). Fuzzy clustering: More than just fuzzification. *Fuzzy Sets and Systems*, 281, 272-279.

166. Zhou, K., Fu, C., and Yang, S. (2014). Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation. *Science China Information Sciences*, 57, 1–8.

167. Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3), 32-57.

168. Höppner, F., and Klawonn, F. (2003). A contribution to convergence theory of fuzzy c-means and derivatives. *IEEE Transactions on Fuzzy Systems*, 11(5), 682-694.

169. Parker, J. K., Hall, L. O., and Bezdek, J. C. (2012). *Comparison of scalable fuzzy clustering methods.* 2012 IEEE International Conference on Fuzzy Systems. Brisbane, Australia, 1-9.

170. Bauer, E., and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1), 105–139.

171. Sill, J., Takacs, G., Mackey, L., and Lin, D. (2009). Feature weighted linear stacking. *Machine Learning*, 37, 1-27.

172. Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2), 241-259.

173. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850.

174. Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10), 1619–1630.

175. Aslam, J. A., Popa, A., and Raluca, A. (2007). On estimating the size and confidence of a statistical audit. *Proceedings of the Electronic Voting Technology Workshop*, 19-20.

176. Schapire, R. E. (1990). The strength of weak learnability, *Machine Learning*, 5, 197-227.

177. Freund, Y., and Schapire, R. E. (1996). *Experiments with a new boosting algorithm*. Thirteenth International Conference on Machine Learning. Shenzhen, China, 148-156.

178. Jaskowiak, P. A., Campello, R. J. G. B. (2011). Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data. *Brazilian Symposium on Bioinformatics*, 1–8.

179. Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont, U.S.A: Lifetime Learning Publications, 149.

180. Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Annals of Statistics*, 40(5), 2733–2763.

181. Platt, J. (1999). Probabilistic outputs for SVMs and comparisons to regularized likelihood methods. In J. Bartlett, B. Scholkopf, D. Schuurmans, and A. J. Smola (Eds.), *Advances in large margin classifiers*. Cambridge: MIT Press, 193.

182. Bordes, A., Ertekin, S., Weston, J., and Bottou, L. (2005). Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6, 1579-1619.

183. Auria, L., and Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis. *Discussion Papers*, 811, 4-19.

184. Han, J., Kamber, M., and Pei, J. (2012). *Data mining: concepts and techniques*. (Third edition). Illinois: Elsevier, 279-281.

185. Shi, B., and Liu, J. (2018). Nonlinear metric learning for kNN and SVMs through geometric transformations. *Neurocomputing*, 318, 18-29.

186. Internet: The Non-Linear Decision Boundary. (2020). URL: https://statinfer.com/204-6-5-the-non-linear-decision-boundary/, Last Accessed: 16 May 2021.

187. Fan, R. E., Chen, P. H., and Lin , C. J. (2005). Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6, 1889–1918.

188. Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1), 77–89.

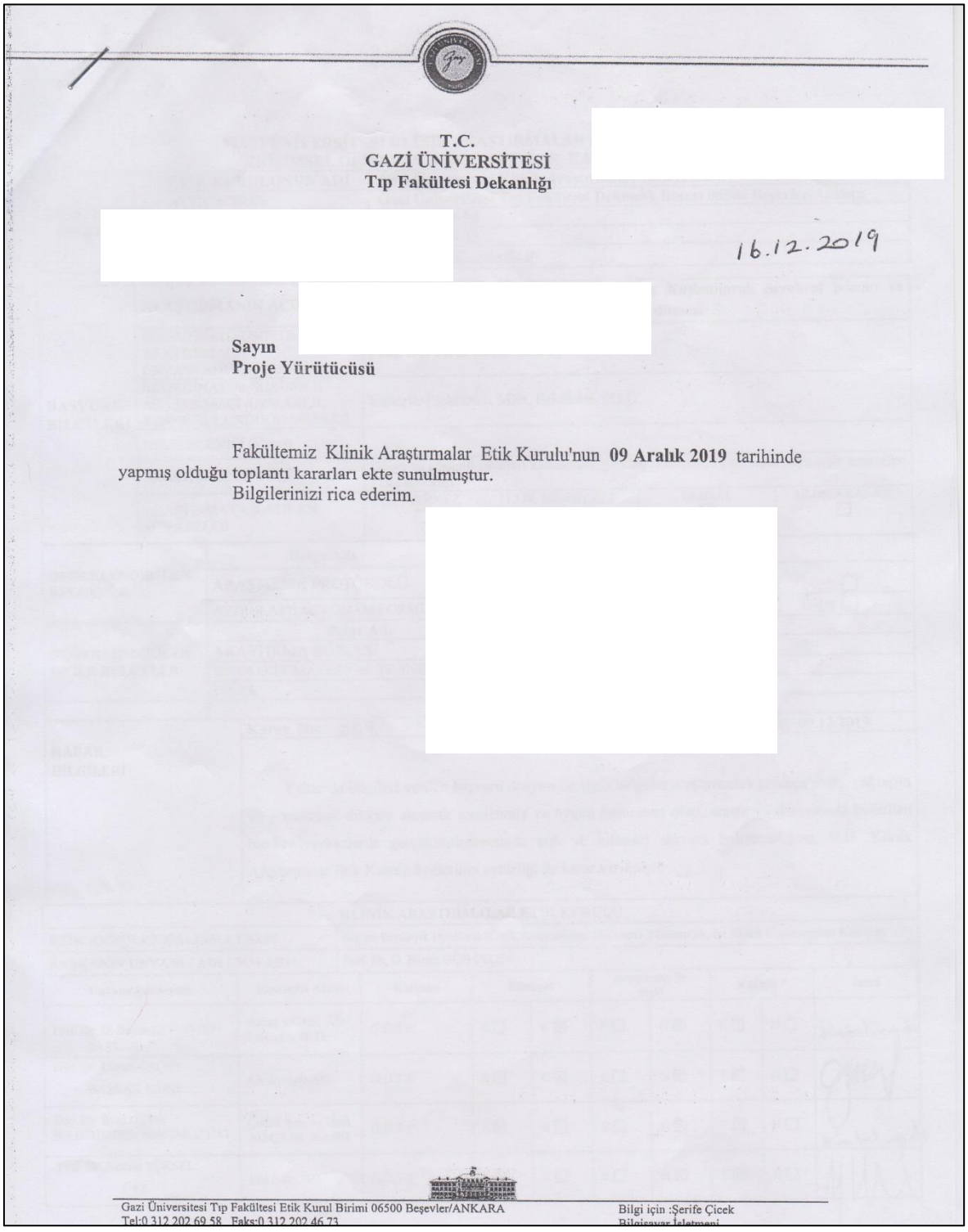
189. Rijsbergen, V. C. J. (1979). *Information retrieval*. (Second Edition). Butterworths: Waltham, 120.

190. Lobo, G. M., Valverde, A. J., and Real, R. (2007). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 1, 1-7.

191. Niewada, M., and Czlonkowska, A. (2014). Prevention of ischemic stroke in clinical practice: a role of internists and general practitioners. *Polskie Archiwum Medycyny Wewnętrznej,* 124(10), 540-548.

APPENDICES

APPENDIX-1. Approval of the clinical research ethics committee



T.C.
GAZİ ÜNİVERSİTESİ
Tıp Fakültesi Dekanlığı

16.12.2019

Sayın
Proje Yürütücüsü

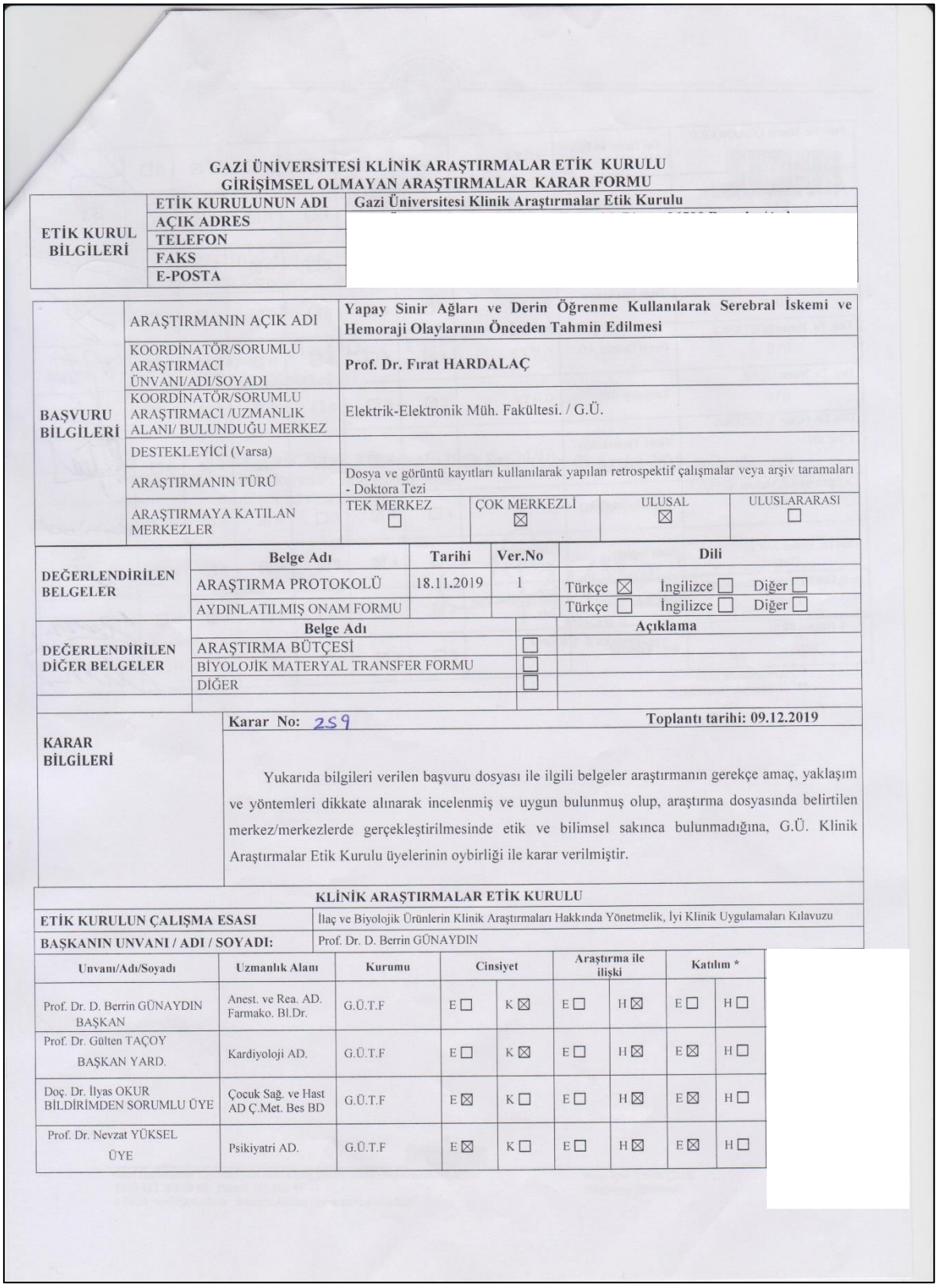
Fakültemiz Klinik Araştırmalar Etik Kurulu'nun **09 Aralık 2019** tarihinde yapmış olduğu toplantı kararları ekte sunulmuştur.
Bilgilerinizi rica ederim.

Gazi Üniversitesi Tıp Fakültesi Etik Kurul Birimi 06500 Beşevler/ANKARA
Tel:0 312 202 69 58 Faks:0 312 202 46 73

Bilgi için :Şerife Çicek
Bilgisayar İşletmeni

Figure 1.1. First page of the approval document

APPENDIX-1. (continued) Approval of the clinical research ethics committee



GAZİ ÜNİVERSİTESİ KLİNİK ARAŞTIRMALAR ETİK KURULU
GİRİŞİMSEL OLMAYAN ARAŞTIRMALAR KARAR FORMU

ETİK KURUL BİLGİLERİ	ETİK KURULUNUN ADI	Gazi Üniversitesi Klinik Araştırmalar Etik Kurulu
	AÇIK ADRES	
	TELEFON	
	FAKS	
	E-POSTA	

BAŞVURU BİLGİLERİ	ARAŞTIRMANIN AÇIK ADI	**Yapay Sinir Ağları ve Derin Öğrenme Kullanılarak Serebral İskemi ve Hemoraji Olaylarının Önceden Tahmin Edilmesi**			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACI ÜNVANI/ADI/SOYADI	**Prof. Dr. Fırat HARDALAÇ**			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACI /UZMANLIK ALANI/ BULUNDUĞU MERKEZ	Elektrik-Elektronik Müh. Fakültesi. / G.Ü.			
	DESTEKLEYİCİ (Varsa)				
	ARAŞTIRMANIN TÜRÜ	Dosya ve görüntü kayıtları kullanılarak yapılan retrospektif çalışmalar veya arşiv taramaları - Doktora Tezi			
	ARAŞTIRMAYA KATILAN MERKEZLER	TEK MERKEZ ☐	ÇOK MERKEZLİ ☒	ULUSAL ☒	ULUSLARARASI ☐

	Belge Adı	Tarihi	Ver.No	Dili
DEĞERLENDİRİLEN BELGELER	ARAŞTIRMA PROTOKOLÜ	18.11.2019	1	Türkçe ☒ İngilizce ☐ Diğer ☐
	AYDINLATILMIŞ ONAM FORMU			Türkçe ☐ İngilizce ☐ Diğer ☐

	Belge Adı		Açıklama
DEĞERLENDİRİLEN DİĞER BELGELER	ARAŞTIRMA BÜTÇESİ	☐	
	BİYOLOJİK MATERYAL TRANSFER FORMU	☐	
	DİĞER	☐	

KARAR BİLGİLERİ

Karar No: 259 **Toplantı tarihi: 09.12.2019**

Yukarıda bilgileri verilen başvuru dosyası ile ilgili belgeler araştırmanın gerekçe amaç, yaklaşım ve yöntemleri dikkate alınarak incelenmiş ve uygun bulunmuş olup, araştırma dosyasında belirtilen merkez/merkezlerde gerçekleştirilmesinde etik ve bilimsel sakınca bulunmadığına, G.Ü. Klinik Araştırmalar Etik Kurulu üyelerinin oybirliği ile karar verilmiştir.

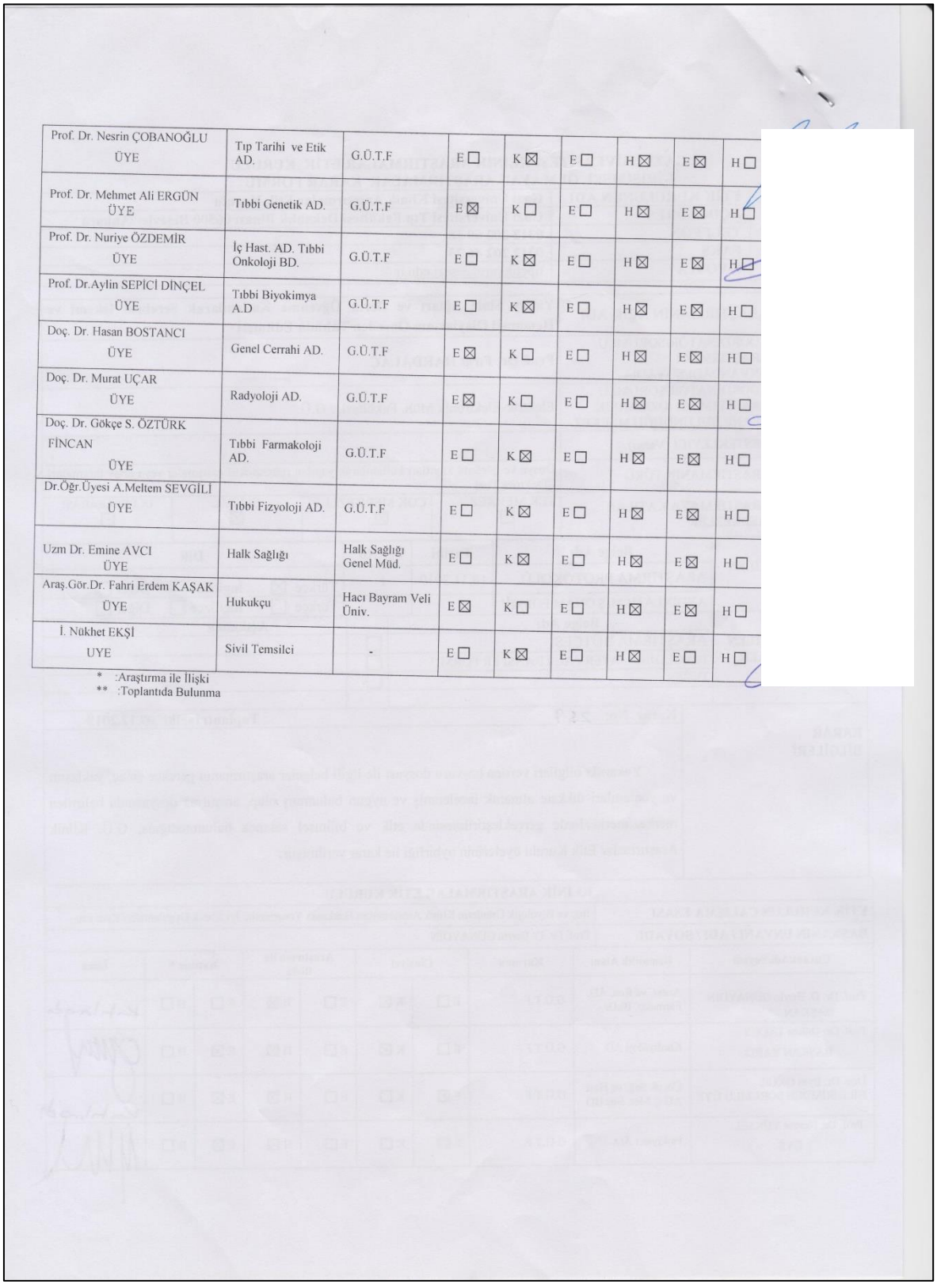
KLİNİK ARAŞTIRMALAR ETİK KURULU

ETİK KURULUN ÇALIŞMA ESASI	İlaç ve Biyolojik Ürünlerin Klinik Araştırmaları Hakkında Yönetmelik, İyi Klinik Uygulamaları Kılavuzu
BAŞKANIN UNVANI / ADI / SOYADI:	Prof. Dr. D. Berrin GÜNAYDIN

Unvanı/Adı/Soyadı	Uzmanlık Alanı	Kurumu	Cinsiyet		Araştırma ile ilişki		Katılım *	
Prof. Dr. D. Berrin GÜNAYDIN BAŞKAN	Anest. ve Rea. AD. Farmako. Bl.Dr.	G.Ü.T.F	E ☐	K ☒	E ☐	H ☒	E ☐	H ☐
Prof. Dr. Gülten TAÇOY BAŞKAN YARD.	Kardiyoloji AD.	G.Ü.T.F	E ☐	K ☒	E ☐	H ☒	E ☒	H ☐
Doç. Dr. İlyas OKUR BİLDİRİMDEN SORUMLU ÜYE	Çocuk Sağ. ve Hast AD Ç.Met. Bes BD	G.Ü.T.F	E ☒	K ☐	E ☐	H ☒	E ☒	H ☐
Prof. Dr. Nevzat YÜKSEL ÜYE	Psikiyatri AD.	G.Ü.T.F	E ☒	K ☐	E ☐	H ☒	E ☒	H ☐

Figure 1.2. Second page of the approval document

APPENDIX-1. (continued) Approval of the clinical research ethics committee



Prof. Dr. Nesrin ÇOBANOĞLU ÜYE	Tıp Tarihi ve Etik AD.	G.Ü.T.F	E☐	K☒	E☐	H☒	E☒	H☐
Prof. Dr. Mehmet Ali ERGÜN ÜYE	Tıbbi Genetik AD.	G.Ü.T.F	E☒	K☐	E☐	H☒	E☒	H☐
Prof. Dr. Nuriye ÖZDEMİR ÜYE	İç Hast. AD. Tıbbi Onkoloji BD.	G.Ü.T.F	E☐	K☒	E☐	H☒	E☒	H☐
Prof. Dr.Aylin SEPİCİ DİNÇEL ÜYE	Tıbbi Biyokimya A.D	G.Ü.T.F	E☐	K☒	E☐	H☒	E☒	H☐
Doç. Dr. Hasan BOSTANCI ÜYE	Genel Cerrahi AD.	G.Ü.T.F	E☒	K☐	E☐	H☒	E☒	H☐
Doç. Dr. Murat UÇAR ÜYE	Radyoloji AD.	G.Ü.T.F	E☒	K☐	E☐	H☒	E☒	H☐
Doç. Dr. Gökçe S. ÖZTÜRK FİNCAN ÜYE	Tıbbi Farmakoloji AD.	G.Ü.T.F	E☐	K☒	E☐	H☒	E☒	H☐
Dr.Öğr.Üyesi A.Meltem SEVGİLİ ÜYE	Tıbbi Fizyoloji AD.	G.Ü.T.F	E☐	K☒	E☐	H☒	E☒	H☐
Uzm Dr. Emine AVCI ÜYE	Halk Sağlığı	Halk Sağlığı Genel Müd.	E☐	K☒	E☐	H☒	E☒	H☐
Araş.Gör.Dr. Fahri Erdem KAŞAK ÜYE	Hukukçu	Hacı Bayram Veli Üniv.	E☒	K☐	E☐	H☒	E☒	H☐
İ. Nükhet EKŞİ UYE	Sivil Temsilci	-	E☐	K☒	E☐	H☒	E☐	H☐

* :Araştırma ile İlişki
** :Toplantıda Bulunma

Figure 1.3. Third page of the approval document

GAZİ GELECEKTİR...