ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD THESIS

Pınar USKANER HEPSAĞ

BIOPSY COST REDUCTION FOR EARLY DIAGNOSIS OF BREAST CANCER USING HYBRID DEEP LEARNING TECHNIQUES

DEPARTMENT OF COMPUTER ENGINEERING

ADANA - 2022

ABSTRACT

Ph.D. THESIS

BIOPSY COST REDUCTION FOR EARLY DISAGNOSIS OF BREAST CANCER USING HYBRID DEEP LEARNING TECHNIQUES

PınarUSKANER HEPSAĞ

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Prof. Dr. Selma Ayşe ÖZEL
Co-supervisor : Prof.Dr. Adnan YAZICI
Year: 2022, Pages: 148
Jury : Prof. Dr. Selma Ayşe ÖZEL
: Prof. Dr. Umut ORHAN
: Assoc. Prof. Dr. Ali İNAN
: Assoc. Prof. Dr. Alper Kamil DEMİR
: Assoc. Prof. Dr. Fatih ABUT

Breast cancer has become one of the most important diseases all over the world. Early diagnosis of breast cancer is very important to reduce the mortality rate. Breast biopsy is one of the most commonly used methods to diagnose breast cancer. However, breast biopsies are sometimes performed even when the patient does not have breast cancer. This leads to various problems for patients such as anxiety, pain, and healthcare costs, etc. Therefore, we proposed a hybrid model to reduce the cost of breast biopsies for early detection of breast cancer using deep learning and machine learning methods. Although many methods have been applied to a range of breast cancer diagnoses using one data type (image or text), the combination of three data types (image, text, and survey) has not been used directly for breast cancer diagnosis in the literature, and the problem of breast cancer diagnosis still needs to be improved. The proposed model combines three types of data, namely image (mammography), text (radiology report), and survey (patient history, physical examination, etc.) of each patient, and generates amalignant or benign output to identify the type of breast cancer. Therefore, a hybrid system consisting of three independent parts is described in this thesis: In the first part, deep learning models such as pre-trained models (VGG16, AlexNet, and ResNet50) and a transformer model are used to classify mammography images. In the second part, machine learning models with different combinations are used to classify surveys. In the third part, similar to images, pre-trained models (BERT versions such as BERTMultilingual, BERTClinical and BERTTurkish) and transformers are used to classify radiology reports. The ensemble model takes the results of each part as input and produces an output to diagnose breast cancer. We also propose a risk-based hybrid neuro fuzzy rule-based system to calculate the risk of breast cancer. We show that using three types of data can reduce the cost of breast biopsies in breast cancer diagnosis. The hybrid model with the hospital dataset improves the precision value by up to 100%.

Key Words: Breast cancer diagnosis, Transformers, VGG16, AlexNet, ResNet50, BERT.

ÖZET

DOKTORA TEZİ

HİBRİT DERİN ÖĞRENME TEKNIKLERI KULLANILARAK MEME KANSERİNDE ERKEN TEŞHİS İÇİN BİYOPSİ MALİYETİNİN DÜŞÜRÜLMESİ

PınarUSKANER HEPSAĞ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Prof. Dr. Selma Ayşe ÖZEL
Eş-danışman : Prof. Dr. Adnan YAZICI
Yıl: 2022, sayfa: 148
Jüri : Prof. Dr. Selma Ayşe ÖZEL
: Prof. Dr. Umut ORHAN
: Doç. Dr. Ali İNAN
: Doç. Dr. Alper Kail DEMİR
: Doç. Dr. Fatih ABUT

Meme kanseri tüm dünyada en önemli hastalıklardan biri haline gelmiştir. Meme kanserinin erken teşhisi, ölüm oranını azaltmak için çok önemlidir. Meme biyopsisi, meme kanserini teşhis etmek için en sık kullanılan yöntemlerden biridir. Ancak bazen meme biyopsileri, hasta meme kanseri olmasa bile yapılır. Bu, hastalar için anksiyete, ağrı ve sağlık maliyetleri vb. gibi çeşitli sorunlara yol açar. Bu nedenle, bu tezdederin öğrenme ve makine öğrenimi yöntemlerini kullanarak meme kanserinin erken teşhisi için meme biyopsilerinin maliyetini azaltmak amacıylahibrit bir model önerilmiştir. Tek tür veri (resim veya metin) kullanılarak meme kanseri teşhisine birçok yöntem uygulanmış olsa da, literatürde meme kanseri teşhisi için üç tür verinin (resim, metin ve anket) kombinasyonu doğrudan kullanılmamıştır ve meme kanseri teşhisi probleminin hala iyileştirilmesi gerekmektedir. Önerilen model, her bir hasta için görüntü (mamogram), metin (radyoloji raporu) ve anket (hasta hikayesi) olmak üzere üç tür veriyi birleştirir ve meme kanserinin türünü belirlemek amacıyla kötü huylu veya iyi huylu bir çıktı üretir. Bu nedenle, bu tezde üç bağımsız parçanın birleşiminden oluşan bir hibrit sistem geliştirilmiştir: Birinci parçada, mamografi görüntülerini sınıflandırmak için önceden eğitilmiş modeller (VGG16, AlexNet ve ResNet50) gibi derin öğrenme modelleri ve bir transformatör modeli kullanılmıştır. İkinci parçada, anketleri sınıflandırmak için farklı kombinasyonlara sahip makine öğrenmesi modelleri kullanılır. Üçüncü parçada, görüntülere benzer şekilde, radyoloji raporlarını sınıflandırmak için önceden eğitilmiş modeller (BERTMultilingual, BERTClinical ve BERTTurkish gibi BERT versiyonları) ve bir transformatör modeli kullanılır. Topluluk modeli, her parçanın sonuçlarını girdi olarak alır ve meme kanseri teşhisi için bir çıktı üretir. Ayrıca, meme kanseri riskini hesaplamak için risk tabanlı hibritnöro+bulanık kural tabanlı bir sistem öneriyoruz. Üç tür verinin kullanılmasının meme kanseri teşhisine meme biyopsisinin maliyetlerini azaltabileceğini göstermekteyiz. Hastane veri setini kullanarak hibrit model, hassasiyet puanını %100'e kadar arttırmaktadır.

Anahtar Kelimeler: Meme kanseri teşhisi, Transformers, VGG16, AlexNet, ResNet50, BERT.

EXTENDED SUMMARY

Among various cancers, breast cancer is the most common cancer in women. It develops when breast cells begin to grow uncontrollably. These cells usually form a tumor that can be seen on a mammogram or ultrasound image or felt as a lump. Breast cancer is divided into two categories: benign and malignant. Benign means that it is not cancer, while malignant is called cancer (Zorluoglu and Agaoglu, 2017). Breast cancer has become one of the most important medical issues due to the increasing number of diseases worldwide (Zand, 2015).

To prevent breast cancer from progressing and leading to death, detection of breast cancer at an early stage is of great importance and requires an accurate diagnostic procedure that allows experts to distinguish malignant from benign disease (Sivakami and Saraswathi, 2015). Ultrasound and mammography are two of the most common radiological procedures for early detection of breast cancer (Sree et al., 2011). However, accurate interpretation of mammograms by radiologists is difficult, resulting in a high number of false positive findings and additional examinations. Due to the high sensitivity but low specificity of mammograms, approximately 20-30% of breast biopsies are detected as malignant (Nolan, 2022). Improving mammography findings to accurately detect breast cancer is very important. This is because a high number of false-positive findings have several negative consequences, such as anxiety, pain, unnecessary surgery, and healthcare costs.

In this thesis, we proposed to develop a novel hbyrid method for breast cancer prediction to reduce the cost of breast biopsies by considering deep learning (DL) models such as Transformers, BERTs, VGG16, and so on. The main contribution in this thesis is that the proposed model combines all three types of data for each patient to reduce the cost of breast biopsies. However, in the literature, only one type of data set has usually been used to classify breast cancer, e.g., only image data, only radiology reports, or only survey data. Another novelty of this thesis is that we collected a new dataset that contains Turkish radiology reports, Turkish surveys and mammograms to

test the performance of the proposed model.Also, in the literature, a survey data has not been used to classify breast cancer.

Mammography method is frequently used by experts to accurately diagnose breast cancer (Ruiz, 2018). Therefore, physicians are increasingly using computer-aided diagnosis methods to interpret mammograms and avoid unnecessary biopsies. Various methods of computational analysis are used, including machine learning algorithms (Sahran et al., 2018). With the increasing success of deep learning methods in computer vision tasks, CNNs have been applied to many image classification problems, especially in medicine (Yadav and Jadhav, 2019). As a result, they have become very popular in breast cancer diagnosis (Alanazi et al., 2021). Deep learning methods require a large dataset to learn better during training. However, medical image datasets are of limited size. Therefore, applying the CNN method to solve medical image analysis problems is a difficult task. Pre-trained models are the most common technique to solve the problem of having limited training data and have been used by researchers to analyze medical images. The most commonly used pre-trained models are ResNet (Chougrad et al., 2017), AlexNet (Hassan et al., 2020) and VGG16 (Falconi et al., 2020) on INbreast, CBIS-DDSM, etc. In this thesis, AlexNet, ResNet50 (He et al., 2016), and VGG16 (Simonyan and Zisserman, 2014) pre-trained models and transformer models (Vaswani et al., 2017) are used to classify our newly created mammography dataset.

Other important data are radiological reports for the prediction of breast cancer. Radiological reports are interpretations of diagnostic imaging, and radiologists write these reports. Radiological reports contain limited vocabulary and are unstructured data (Suárez-Paniagua et al., 2021). Therefore, it is very difficult to extract meaningful data from reports automatically. On the other hand, extracting information from these reports is very important for early detection of breast cancer. BERT and Transformer models are used to analyze our newly created Turkish radiology report dataset.

A survey to predict breast cancer is also used in this work. A survey of patients with breast symptoms is conducted. Different machine learning algorithms are used to classify our newly created dataset of Turkish surveys.

The best models whose performance was highest for each data type are selected, and the proposed model receives the results of the most successful models as input. Then, the proposedhybridsystem provides a result of whether the patient is malignant or benign to reduce the cost of breast biopsies. In addition, we used the combination of text and image data with the Transformer model to determine whether breast cancer prediction can be improved by the late fusion method.

One of the problems in breast cancer detection is having imbalanced data, which affects the quality of training. Therefore, data augmentation methods were used to increase the number of samples in the data set and obtain the same number of samples in the two classes. Since the number of radiology reports, mammogram images, and surveys are small and unbalanced, EDA expansion and self-copy for radiology reports, simple transformations for mammogram images and self-copy for surveys, and weighted random sampling methods are used for this purpose.

The experimental results show that the pre-trained VGG16 model performs the best for mammograms, the Transformer model with $BERT_{multilingual}$ performs the best for radiology reports, and Attribute selectedclassifier with pure data performs the best for surveys. The experimental results also reveal that the hybrid model outperformsall other models in predicting patient data with benign tumors as benign. In other words, the proposed model achieves a precision score of nearly 100% for our collected dataset. Thus, our hybrid model reduces the cost of breast biopsies. Therefore, we can say that our proposed model is the best choice for breast cancer detection using newly collected Turkish radiology reports, newly collectedTurkish surveys, and newly collected mammograms.

In this thesis, we also proposed a system which is called a risk-based hybrid neuro+fuzzy rule-based to compute breast cancer risk score from mammograms. This proposed model calculates breast cancer risk using a combination of DL (CNN-base and Resnet50) and fuzzy rule-based system (FRBS) to protect people from breast

cancer. Moreover, this FRBS creates a simple rule-based classifier to diagnose breast cancer based on these risk values. In this proposedsystem, the deep learning part works as a preprocessor. It generates probabilities for each class from the training data for fuzzy rules or membership functions and then fades into the background. Then the fuzzy system and the classifier come into play, working as postprocessors in the hybrid model. The othercontribution in this thesis is that the proposed model captures uncertain data well and expresses them in real numbers to reduce unneeded biopsies and the associated negative consequences such as health care costs, pain, anxiety, and so onin diagnosing breast cancer. However, in the literature, there is no such study to classify mammogram images.

The experimental results show that a combination of Deep neural network (DNN) with FRBS leads to better classification results for breast cancer diagnosis than a pure CNN and a pure ResNet50.The proposed model improves the precision score from 87% to 91% as well as improving accuracy and F1-score compared to pure deep learning model in diagnosing breast cancer using mammograms. As a result, the number of breast biopsies can be reduced and breast cancer can be diagnosed better withour hybrid neuro+fuzzymodel.

On the other hand, in this thesis, besides analyzing different combinations such as text-image combination and image-image combination by late fusion, we also use a hybrid model of pre-trained + Transformer for mammogram classification.

We believe that this thesis will serve as a guide that will provide useful tips to researchers in the field of breast cancer detection for selecting the best models for classifying mammogram images, surveys, and radiology reports in order to develop more effective and reliable studies. In the future, we plan to evaluate the performance of the proposed model with different learning models. We also plan to use Vision Transformer to improve the results of mammogram images for Transformers.

GENİŞLETİLMİŞ ÖZET

Çeşitli kanser türleri arasında meme kanseri kadınlarda en sık görülen kanserdir. Meme hücreleri kontrolsüz bir şekilde büyümeye başladığında gelişir. Bu hücreler genellikle bir mamogram veya ultrason görüntüsünde görülebilen veya bir yumru olarak hissedilebilen bir tümör oluşturur. Meme kanseri iki kategoriye ayrılır: iyi huylu ve kötü huylu. İyi huylu kanser olmadığı anlamına gelirken, kötü huylu ise kanser olarak adlandırılmaktadır (Zorluoğlu ve Ağaoğlu, 2017). Meme kanseri, dünya çapında artan hastalık sayısı nedeniyle en önemli tıbbi sorunlardan biri haline gelmiştir (Zand, 2015).

Meme kanserinin ilerlemesini ve ölüme yol açmasını önlemek için, meme kanserinin erken bir aşamada saptanması büyük önem taşır ve uzmanların kötü huylu hastalığı iyi huylu hastalıktan ayırt etmesine olanak tanıyan doğru bir teşhis prosedürü gerektirir (Sivakami ve Saraswathi, 2015). Ultrason ve mamografi, meme kanserinin erken teşhisi için en yaygın radyolojik prosedürlerden ikisidir (Sree ve ark., 2011). Ancak, radyologlar tarafından mamogramların doğru yorumlanması zordur, bu da çok sayıda yanlış pozitif bulguya ve ek incelemelere neden olur. Mamogramların duyarlılığının yüksek ancak özgüllüğünün düşük olması nedeniyle meme biyopsilerinin yaklaşık %20-30'u malign olarak saptanır (Nolan, 2022). Meme kanserini doğru bir şekilde tespit etmek için mamografi bulgularının iyileştirilmesi çok önemlidir. Bunun nedeni, çok sayıda yanlış pozitif bulgunun kaygı, ağrı, gereksiz cerrahi ve sağlık maliyetleri gibi çeşitli olumsuz sonuçları olmasıdır.

Bu tezde, Transformers, BERTs, VGG16 vb. derin öğrenme modellerini göz önünde bulundurarak meme biyopsilerinin maliyetini azaltmak amacıyla meme kanseri tahmini için yeni bir hibrit yöntem önerilmektedir.

Bu tezdeki ana yenilik, önerilen modelin meme biyopsilerinin maliyetini azaltmak amacıyla her hasta için üç tür veriyi bir araya getirmesidir. Bununla birlikte, literatürde, meme kanserini sınıflandırmak için genellikle yalnızca bir tür veri seti kullanılmıştır; örneğin, yalnızca görüntü verileri, yalnızca radyoloji raporları veya yalnızca anket verileri. Bu tezin bir diğer yeniliği, önerilen modelin performansını test etmek için Türkçe radyoloji raporları, Türkçe anketler ve mamogramlardanoluşanyeni

bir veri kümesi toplamamızdır. Ayrıca literatürde meme kanserini sınıflandırmak için bir anket verisi kullanılmamıştır.

Mamografi yöntemi, uzmanlar tarafından meme kanserini doğru bir şekilde teşhis etmek için sıklıkla kullanılmaktadır (Ruiz, 2018). Bu nedenle doktorlar, mamogramları yorumlamak ve gereksiz biyopsilerden kaçınmak için bilgisayar destekli tanı yöntemlerini giderek daha fazla kullanıyor. Bu amaçla, makine öğrenimi algoritmaları da dâhil olmak üzere çeşitli hesaplamalı analiz yöntemleri kullanılmaktadır (Sahran ve diğerleri, 2018). Bilgisayarlı görü görevlerinde derin öğrenme yöntemlerinin artan başarısı ile birlikte CNN'ler başta tıp olmak üzere birçok görüntü sınıflandırma problemine uygulanmıştır (Yadav ve Jadhav, 2019)vebu yöntemler meme kanseri tanısında oldukça popüler hale gelmiştir (Alanazi ve ark., 2021). Derin öğrenme yöntemleri, eğitim sırasında daha iyi öğrenmek için büyük bir veri kümesi gerektirir. Bununla birlikte, tıbbi görüntü veri kümeleri sınırlı boyuttadır. Bu nedenle az sayıdaki tıbbi görüntülerden öğrenme problemlerini çözmek için CNN yöntemini uygulamak zor bir iştir. Önceden eğitilmiş modeller, sınırlı sayıdaki eğitim verileri sorununu çözmek için en yaygın uygulanan tekniktir ve araştırmacılar tarafından tıbbi görüntüleri analiz etmek amacıyla kullanılmıştır. En yaygın olarak kullanılan önceden eğitilmiş modeller ResNet (Chougrad ve diğerleri, 2017), AlexNet (Hassan ve diğerleri, 2020) ve INbreast, CBIS-DDSM, vb. üzerinde VGG16 (Falconi ve diğerleri, 2020) olarak sıralanabilir. Bu tezde, AlexNet, ResNet50 (He ve ark., 2016) ve VGG16 (Simonyan ve Zisserman, 2014) önceden eğitilmiş modeller ve Transformermodelleri (Vaswani ve ark., 2017) yeni oluşturulan mamografi veri setimizi sınıflandırmak için kullanılmıştır.

Meme kanseri teşhisine yönelik diğer önemli verikaynağıise radyolojik raporlardır. Radyolojik raporlar, tanısal görüntülemenin yorumlarıdır ve radyologlar bu raporları yazar. Radyolojik raporlar sınırlı sayıda farklı kelime içerir ve yapılandırılmamış verilerdir (Suárez-Paniagua ve diğerleri, 2021). Bu nedenle, raporlardan otomatik olarak anlamlı veriler çıkarmak çok zordur. Öte yandan bu raporlardan bilgi çıkarmak meme kanserinin erken teşhisi için çok önemlidir. Bu tezde, yeni oluşturulan Türkçe radyoloji raporu veri setimizi analiz etmek için BERT ve Tansformer modelleri kullanılmıştır.

Bu çalışmada meme kanserini teşhis etmeye yönelik bir anket de kullanılmaktadır. Meme semptomları olan hastalarda bir anket yapılır. Yeni oluşturulan Türkçe anket veri setimizi sınıflandırmak için farklı makine öğrenme algoritmaları kullanılmıştır.

Bu tez kapsamında, her bir veri türü için sınıflama performansı en yüksek olan en iyi modeller seçilip bu modellerin sonuçları hibrit modele girdi olarak kullanılır. Daha sonra hibritmodelimiz, meme biyopsilerinin maliyetini azaltmak için hastanın durumunun kötü huylu mu yoksa iyi huylu mu olduğu sonucunu verir. Ek olarak, meme kanseri teşhisinin geç füzyon yöntemiyle iyileştirilip iyileştirilemeyeceğini belirlemek için metin ve görüntü verilerinin Transformer modeliyle birleşimi kullanılmıştır.

Meme kanseri teşhisindeki sorunlardan biri de eğitimin kalitesini etkileyen dengesiz dağılımlı verilerdir. Bu nedenle, veri setini artırmak ve iki sınıfta aynı sayıda örnek elde etmek için veri artırma yöntemleri kullanılmıştır. Radyoloji raporlarının, mamogram görüntülerinin ve anketlerin sayısı az ve dengesiz olduğundan, radyoloji raporları için EDA genişletme ve kopyalama, mamogram görüntüleri için basit dönüşümler ve anketler için kopyalama ve ağırlıklı rastgele örnekleme yöntemleri kullanılmaktadır.

Deneysel sonuçlar, önceden eğitilmiş VGG16 modelinin mamogramlar için en iyi performansı gösterdiğini, $BERT_{multilingual}$ özelliğine sahip Transformer modelinin radyoloji raporları için en iyi performansı verdiğini ve saf verilerle nitelik seçimi sınıflayıcı yönteminin anketler için en iyi performansa sahip olduğunu göstermektedir. Deneysel sonuçlar, hibritmodeliniyi huylu tümöre sahip hastaverileriniköt huylu tümör yerine iyi huylu tümör olarak tahmin etmede diğer tüm modellerden daha iyi performans gösterdiğini ortaya koymaktadır. Başka bir deyişle, önerilen model, toplanan veri kümemiz için yaklaşık %100'lük bir kesinlik puanı elde etmektedir. Böylece hibrit modelimiz meme biyopsilerinin maliyetini düşürmek amacıyla kullanılabilir. Bu nedenle, önerilen modelimizin bu tez kapsamında derlenen Türkçe radyoloji raporları, Türkçe anketler ve mamogramlardan oluşan meme kanseri veri kümesi için en iyi seçim olduğunu söyleyebiliriz.

Bu tezde, mamogramlardanmeme kanseri risk puanını hesaplamak için risk tabanlı hibritnöro+bulanık kural tabanlı olarak adlandırılan bir sistem önerdik. Önerilen

bu model, insanları meme kanserinden korumak için derin öğrenme (CNN tabanlı ve Resnet50) ve bulanık kural tabanlı sistemin bir kombinasyonunu kullanarak meme kanseri riskini hesaplar. Ayrıca, bu bulanık kural tabanlı sistem, bu risk değerlerine dayalı olarak meme kanserini teşhis etmek için basit bir kural tabanlı sınıflandırıcı oluşturur. Önerilen bu sistemde, derin öğrenme kısmı bir önişlemci olarak çalışmaktadır. Bulanık kurallar veya üyelik fonksiyonları için eğitim verilerinden her sınıf için olasılıklar üretir ve ardından arka plana geçer. Daha sonra hibrit modelde son işlemci olarak çalışan bulanık sistem ve sınıflandırıcı devreye girer.

Bu tezdeki diğer katkı, önerilen modelin belirsiz verileri iyi bir şekilde yakalaması ve meme kanseri teşhisinde gereksiz biyopsileri ve maliyet, ağrı, kaygı vb. ilişkili olumsuz sonuçları azaltmak için bunları gerçek sayılarla ifade etmesidir. Ancak literatürde mamografi görüntülerini sınıflandıran böyle bir çalışma bulunmamaktadır.

Deneysel sonuçlar, derin sinir ağı ile bulanık kural tabanlı sistem kombinasyonunun meme kanseri teşhisi için saf CNN ve saf ResNet50'den daha iyi sınıflandırma sonuçlarına sahip olduğunu göstermektedir. Önerilen model, mamogramlar kullanılarak meme kanseri teşhisinde saf derin öğrenme modeline kıyasla kesinlik puanını %87'den %91'e çıkarmanın yanı sıra doğruluğu ve F1 puanını iyileştirmektedir. Sonuç olarak, hibritnöro+bulanık modelimiz ile meme biyopsisi sayısı azaltılabilir ve meme kanseri daha iyi teşhis edilebilir.

Öte yandan, bu tezde, metin-görüntü kombinasyonu ve görüntü-görüntü kombinasyonu gibi farklı kombinasyonları geç füzyonla analiz etmenin yanı sıra, mamogram sınıflandırması için önceden eğitilmiş + Transformerhibrit modelini de kullanıyoruz.

Bu tezin, meme kanseri tespiti alanındaki araştırmacılara, mamogram görüntülerini, anketleri ve radyoloji raporlarını sınıflandırmak için en iyi modelleri seçme konusunda daha etkili ve güvenilir çalışmalar geliştirmek için faydalı ipuçları sunacak bir rehber olacağına inanıyoruz. Gelecekte, önerilen modelin performansını farklı öğrenme modelleri ile değerlendirmeyi planlıyoruz. Ayrıca mamogram görüntülerinin Transformer model sonuçlarını iyileştirmek için VisionTransformer kullanmayı planlıyoruz.

ACKNOWLEDGEMENTS

I would like to present my endless gratitude to my supervisors Prof. Dr. Selma Ayşe ÖZEL and Prof. Dr. Adnan YAZICI for their unlimited support, guidance, patience and motivation for the successful completion of this research work.

I wish to express my special thanks to Prof. Dr.Umut ORHAN, Assoc. Prof. Dr. Ali İNAN, and Asst. Prof. Dr. Kubilay DALCI for their valuable help, support and suggestions.I would like to thank Assoc. Prof. Dr. Alper Kamil DEMİR and Assoc. Prof. Dr. Fatih ABUT for being part of the jury for the defense of my thesis.

I am also thankful to Umut HEPSAĞ and Necva BÖLÜCÜ for their support and motivation.

CONTENTS **PAGE**

LIST OF TABLES ..... PAGE

LIST OF FIGURES **PAGE**

LIST OF ABBREVIATIONS

ADs : Architectural Distortions
AI : Artificial Intelligence
ANN : Artificial Neural Network
ANFIS : Adaptive Neuro Fuzzy Inferences System
BC : Breast Cancer
BCDR-D02 : Breast Cancer Digital Repository
BERT : Bidirectional Encoder Representations from Transformers
BI-RADS :Breast Imaging-Reporting and Data System
BROK : BI-RADS Observation Kit
CAD : Computer Aided Diagnosis
CBIS : Curated Breast Imaging Subset
CFS : Correlation-based Feature Selection
CNN : Convolutional Neural Networks
CT : Computed Tomography
DCE-MRI : Dynamic Contrast Enhanced Magnetic Resonance Imaging
DDSM : Digital Database for Screening Mammography
DFS : Distinguishing Feature Selector
DICOM : Digital Imaging and Communications in Medicine
DL : Deep Learning
DT : Decision Tree
EDA : Easy Data Augmentation
eICU : Electronic Intensive Care Unit
FFN : Feed Forward Network
FL : Fuzzy Logic
FIS : Fuzzy Inference System
ICD : International Classification of Diseases
IT : Information Technology

KNN : K-Nearest Neighbor
LR : Logistic Regression
LSTM : Long Short Term Memory
MCs : Microcalcifications
MHA : Multi-Head Attention
MIAS : Mammographic Image Analysis Society
MIMIC : Medical Information Mart for Intensive Care
ML : Machine Learning
MLF : Multilevel Fuzzy
MLM : Masked Language Model
MLP : Multi-Layer Perceptron
MNB : Multinomial Naïve Bayes
NB : Naive Bayesian
NCI : National Cancer Institute
NLP : Natural Language Processing
NSP : Next Sentence Prediction
PACS : Picture Archiving and Communication Systems
PCA : Principle Component Analysis
PRISMA : Preferred Reporting Items for Systematic Reviews and Meta-Analysis
ReLU : Rectified Linear Unit
ResNet : Residual Neural Network
RF : Random Forest
RNN : Recurrent Neural Network
SGD : Stochastic Gradient Descent
SMOTE : Synthetic Minority Oversampling Technique
SVM : Support Vector Machine
TF-IDF : Term Frequency – Inverse Document Frequency
UCI : University of California Irvine
VGG : Visual Geometry Group

WBC : Wisconsin Breast Cancer

WDBC : Wisconsin Diagnostic Breast Cancer

WEKA : Waikato Environment of Knowledge Analysis

WPBC : Wisconsin Prognostic Breast Cancer

WRS : Weighted Random Sampling

XGBoost : Extreme Gradient Boosting

1. INTRODUCTION

Cancer is the second leading cause of death worldwide, responsible for an estimated 9.6 million deaths in 2018 (Cancer, 2022). The estimated 2.3 million new cases suggest that by 2020, one in eight cancers diagnosed will be breast cancer (Arafat et al., 2021; Breast Cancer, 2021). Breast cancer, which is an uncontrolled growth of breast cells, is the most commonly diagnosed cancer in women worldwide, accounting for one in four cancer cases (Ahmedet al., 2021). Tumors arising from uncontrolled growth of breast cells are divided into two classes. One is benign and is referred to as noncancerous, and the other is malignant and is referred to as cancerous (Cancer, 2022; Ahmed et al., 2021).

Breast cancer (BC) is a complex disease that may be the result of an interaction between genetic factors, environment, lifestyle, and hormones that contribute to the growth of abnormal cells in the breast. There are modifiable and non-modifiable environmental factors that increase the incidence of breast cancer. Non-modifiable factors include gender, age, genetic characteristics, including family or personal history of breast cancer, ethnicity, and early menarche or menopause. Modifiable risk factors, usually related to lifestyle, include alcohol consumption, overweight or obesity, physical inactivity, parity, and use of certain medications such as oral contraceptives (Kapoor et al., 2020; Youn& Han, 2020). As a woman ages, her risk of developing breast cancer increases due to genetic and hormonal factors (Cancer, 2022; Sun et al., 2017). According to studies, most cases are diagnosed after the age of 50. Other risk factors or causes for breast cancer occurrence include early age of menarche or late menopause. Women who do not have children or have their first child after age 30 are also at increased risk (Sun et al., 2017).

Early detection of breast cancer is vital for the patient, because late detected breast cancer often leads to the death of the patient. When breast cancer is diagnosed early, both treatments are less painful for the patient and the risk of

death is lower. Nevertheless, a correct and valid diagnostic method is necessary to detect BC at an early stage (Irvin et al., 2020).

Mammography is one of the most common radiological methods for early detection of breast cancer. It is difficult to correctly diagnose BC from mammograms because mammogram images have both high sensitivity and low specificity (Prummel et al., 2016). This can make it appear that the person has BC, even if he or she does not. For this reason, a method is needed to prevent the person from having BC even if he or she does not, or to prevent the person from not having BC even if he or she does, when interpreting mammography results to diagnose BC. Nowadays, the use of Computer Aided Diagnosis (CAD) in the diagnosis of BC is very common among physicians. The reason is that doctors can better interpret mammograms, which prevents the person from having BC even if they do not have it (Dromain et al., 2013).

There are a number of CAD methods for radiological image data (Goreke et al., 2016) and Deep Learning (DL) is the best CAD method for analyzing images (Uzunhisarcıklı et al., 2020). The use of DL is widely applied in BC classification (Shen, 2017; Lévy and Jain, 2016). Nevertheless, there are a number of studies that classify mammograms according to the results of DL to determine if a person has BC(Jadoon, Zhang, Haq, Butt, Jadoon, 2017; Hepsağ et al., 2017).

Mammography reports are another source of data to extract knowledge from mammograms and diagnose breast cancer at an early stage. Radiologists attempt to identify critical features such as Microcalcifications (MCs), Architectural Distortions (ADs), and biomarkers of cancer risk (Abdelrahman et al., 2021). Radiologic reports are interpretations of diagnostic imaging, and a radiologist writes these reports. Radiologists typically use complex medical terms and limited vocabulary when writing reports. Therefore, radiology reports contain a smaller vocabulary than other electronic health records (Suárez-Paniagua et al., 2021). Since radiology reports are unstructured, it is very difficult to extract meaningful data automatically from them. On the other hand, extracting

information from radiology reports is very important to help physicians make more accurate decisions when diagnosing breast cancer (Casey et al., 2021).

The process of data collection, in which data such as medical history, physical examination, or risk factors are collected in addition to a person's demographic information, is called a survey. In the data collection process, we created and collected surveys in addition to the mammography images and radiology report to obtain information about each patient's medical history, lifestyle, etc. to help physicians make clinical decisions to diagnose breast cancer.

In this thesis, we proposed a system which is called a risk-based hybrid neuro+fuzzy rule-based to compute BC risk scorefrom mammograms. This proposed model calculates BC risk using a combination of DL (CNN-base and Resnet50) and fuzzy rule-based system (FRBS) to protect people from BC. Moreover, this FRBS creates a simple rule-based classifier to diagnose BC based on these risk values.The proposed system consists of two independent parts: In the first part, the mammogram is given as input to the DL model to generate probabilities for the “malignant” and “benign” types, which serve as input to the fuzzy system. In the second part, these probabilities are used by FRBS to calculate the BC risk score. In addition, the proposed model uses a simple rule-based classifier to decide on the labelling of the mammograms, i.e., whether it is a "benign" or "malignant" type, based on the risk scores.

In our thesis, we have also developed a model to decide whether the mammograms are classified as “benign” or “malignant” based on three types of data such as text, image and survey from each patient. We called this model a hybrid model because it uses three different types of data. Our hybrid model takes each patient's mammogram image, survey, and radiology report as input and makes a prediction for each patient as to whether she belongs to the malignant or benign classes.

To work with these text, image, and survey data of each patient, we applied different analysis methods for each data type and selected the best models for each

data type. The labels predicted by the best models for each patient data type are then used as input to the ensemble learning model to predict the breast cancer class for each patient. Another novelty of this thesis is that we collected a new dataset that contains Turkish radiology reports, Turkish surveys and mammograms to test the performance of the proposed model.

1.1. The Aims and Objectives of this Thesis

The main objective of this thesis is to contribute to previous studies by developing a hybrid system that can help determine biopsy operations by considering deep learning approaches using patient data such as surveys, radiological reports, and image data. The second objective of the thesis is to collect a dataset to make performance evaluation of the proposed methods.

The performance of the hybrid model was compared with state-of-the-art methods using our collected dataset from the Department of General Surgery at Çukurova University.

1.2. Contributions of this Thesis

There are many studies in the literature to improve the accuracy of breast cancer classification using only image data, only report data, or only survey data. In this thesis, a hybrid model is developed and applied to all three types of data: Image, Text and Survey, to show how it affects the prediction of breast cancer. Fuzzy logic is one of the most used methods for BC diagnosis. In this thesis, a risk-based hybrid neuro+fuzzy rule-based system is designed to compute breast cancer risk score and using these risks for BC diagnosis. This fuzzy approach shows that capturing uncertain data improves the classification of breast cancer. The main contribution of this work can be summarised as follows:

- Only one type of data is usually collected in the literature, while we created a new dataset from radiology reports, mammography images, and surveys obtained from real breast cancer patients.

- In this thesis, we developed a risk-based hybrid neuro+fuzzy rule-based system for BC risk prediction and BC diagnosis using mammograms. Our hybrid system calculates risk scores of BC using neuro+fuzzy system and assigns class labels using the calculated risk scores with a rule-based classifier. To our knowledge, there is no study using a deep neural network (DNN) architecture that combines neural network-generated probabilities with a rule-based fuzzy system to improve BC diagnosis.

- We created fuzzy rules for determining BC risk and developed a rule-based classifier to decide on mammogram labelling based on risk scores.

- In this thesis, we designed a hybrid system for breast cancer diagnosis using a combination of mammogram images, surveys, and radiological reports from individual patients. To our knowledge, this is the first study to use three types of data from each patient simultaneously to classify breast calcifications as malignant or benign.

- We also created and collected a survey to obtain patients' data on medical history, physical examinations, radiological findings, etc. for breast cancer detection. To our knowledge, this is the first study to use survey data for breast cancer classification.

- We compared the effects of preprocessing techniques such as augmentation and balancing on classification performance using the newly collected patient datasets consisting of mammography images, radiology reports, and surveys.

- A comprehensive analysis of machine learning models including deep learning techniquesis presented for each data type.

- Our study can also be used as a guide that provides useful tips for selecting the best classification models for mammography images, surveys, and radiology reports for researchers in breast cancer detection to develop more effective and reliable studies.

1.3. Outline of the Thesis

This thesis is organized as follows:

In chapter two, the literature has been reviewed and important studies that have been addressed for survey, text, and image classification techniques for classifying breast cancer as malignant or benign are explained.

In chapter three, the materials and methods are explained. All the other materials that were used to conduct the study are explained individually.

In chapter four, the results of the proposed method are analyzed in detail. In this chapter, the evaluation metrics and experimental setup for the task are summarized and our experimental results are presented along with the data set.

The last chapter, chapter five, contains the conclusion for this thesis; it includes comments and observations about the study and future prospects for the implementation of the proposed method.

2. LITERATURE REVIEW

The method proposed in this thesis consists of classifying 3 kinds of datasets: questionnaires (surveys), images, and reports. Since each data typewas analysed in detail for breast cancer classification, the literature review of each data typewas given separately. Eachdatatypeintroduces the main approaches used in machine learning methods and classification methods for breast cancer. Besides, we also summarize the latest deep learning techniques in breast cancer classification.

2.1. Survey Classification Approaches

In the survey data of the thesis work, we intend to test the performance of different machine learning, deep learning and ensemble learning algorithms for detecting breast cancer. We focus on the different algorithms and compare the classification results to find the most accurate algorithm for the breast cancer classification problem. For this purpose, we reviewed the previous studies on breast cancer prediction algorithms based on Support Vector Machine (SVM), J48, Naïve Bayes (NB), Adaboost, Bagging, Deep Learning 4J (DL4J), BayesNet (BN), DecisionStump (DS), AttributeSelectedClassifier, and RandomForest (RF) classifiers (Fatima et al., 2020). Most of the studies in the literature used these classifiers with the features obtained from mammography images and made breast cancer predictions. In our work, we use these classifiers with the features obtained from patient history, physical examination as well as features obtained from mammography images.

Fakoor et al. (2013), Ramadevi et al. (2015) and Israni (2019) used Principal Component Analysis (PCA) to solve the dimensionality problem. It is well known that dimensionality reduction is very important to improve classification accuracy and speed up training. Fakoor et al. (2013) used the unsupervised and deep learning methods to improve cancer diagnosis. The

proposed method (Fakoor et al., 2013) using PCA derived sparse autoencoder features and randomly selected raw features outperforms the unsupervised method and improves the accuracy in cancer classification. This study differs from others using autoencoders to classify breast cancer. On the other hand, Ramadevi et al. (2015) used a dimensionality reduction method to classify multiple breast cancer datasets which are Wisconsin Prognostic Breast Cancer (WPBC), Wisconsin Diagnostic Breast Cancer (WDBC), etc.. Each breast cancer dataset is classified using K-Nearest Neighbors (KNN), SVM, C4.5 decision tree (C4.5), Logistic Regression (LR) and RF classification methods. The experimental results show that after applying the PCA, the performance of the classifiers has increased on some datasets; on some datasets it has decreased and on some it has remained the same.For the tests of breast cancer data, Israni, (2019) used PCA to reduce dimension for the purpose of feature selection and extractionand compared Tree, SVM, KNN, Stochastic Gradient Decent (SGD), RF, NN, NB, and AdaBoost algorithms. In this study, the Wisconsin Diagnostic Breast Cancer (WDBC) dataset was used. SVM showed the highest accuracy of 92.78%.

Studies Christobel and Sivaprakasam (2011), Khourdifi and Bahaj (2018), Maysanjaya et al. (2018), and Maliha et al. (2019), use NB classifier for BC prediction. Christobel and Sivaprakasam (2011) compared C4.5, NB, SVM and KNN, to find out the most accurate algorithm. In this study, SVM outperformed all other algorithms with 96.99% accuracy in classifying breast cancer. Khourdifi and Bahaj (2018) compare RF, SVM and KNN as well as NB for breast cancer prediction. In this study, the authors focused on bioinformatics and medical science classification techniques and compared data mining algorithms to find the most accurate algorithm for breast cancer prediction. Similar to Christobel and Sivaprakasam (2011), this study showed that SVM was the best classifier for analyzing WDBC with 97.9% accuracy. Maysanjaya et al. (2018) used a combination of wrapper and NB algorithms to improve the classification results of breast cancer datasets. The authors used the wrapper algorithm for feature selection

and the NB algorithm for the classification process. This study differs from others using wrapper to select features. It was found that the combination of wrapper and NB improved the accuracy of breast cancer diagnosis by up to 99%. Research in Maliha et al. (2019) compares NB, KNN, and J48 classifiers for predicting 9 different types of cancer such as breast cancer, brain cancer, blood cancer, ovarian cancer, prostate cancer etc. In this paper, data collected with the help of doctors and experts has been used. The collected data includes 1059 instances and 61 characteristics. Based on the test results, the symptoms were checked. If they show the same result, it means that the test result is correct. When comparing the results of the three breast cancer detection algorithms, J48 performed the worst with 98.5% accuracy and KNN performed the best with 98.8% accuracy. This study differs from others using their collected data while Maysanjaya et al. (2018) and Khourdifi and Bahaj (2018) using WDBC dataset.

Keleş (2019) used Bagging, Instance based learning with some parameters, Random Committee (RC), RF, Simple Classification and Regression Tree (RT) algorithms for breast cancer diagnosis. The WEKA tool was used to analyze the Antenna dataset. The authors found that RF algorithm has the best accuracy in detecting breast cancer. The second best results with 90.9% accuracy were obtained by the Bagging and RC algorithms.

Breast cancer prediction was performed by the authors using Artificial Neural Networks (ANNs) and SVM for Wisconsin Breast Cancer (WBC) in (Bayrak et al., 2019). The authors used K-fold cross validation for each technique. The performance of the two techniques was tested using the tool WEKA. The comparison of the results showed that SVM performed better than ANNs with 99.6% accuracy. Similar to Fakoor et al. (2013), authors used deep learning for classification BC.

Islam et al. (2020) focuses on the comparison of five supervised machine learning techniques, namely SVM, KNN, RF, ANNs and LR, for breast cancer detection. This study is similar to Khourdifi and Bahaj (2018) in that it uses the

same algorithms except ANNs and LR to classify the WDBC dataset. They report a comparative analysis between theclassification techniques using the WDBC. Their experimental results obtain the highest accuracy of 98.57% by ANNs, while the lowest accuracy is obtained by RF and LR with an accuracy of 95.7%.

A summary of previous studies on breast cancer classification is provided in Table 2.1. As can be seen from the table, the studies generally use a dataset containing characteristic features of images. None of these studies use survey data to apply machine learning algorithms for BC classification.

In our thesis, we created surveys with the help of a general surgeon and collected a survey dataset by applying patients who came to the General Surgery department with breast complaints. Creating and collecting survey dataset makes our study different from others. To the best of our knowledge, there is no study, which creates and collects surveys for breast cancer classification. We use machine learning methods for analyzing survey data because our newly created survey data is similar to the dataset used in previous studies. In this work, survey data are classified using different combinations of the different augmentation, normalization, and feature selection techniques to find the most successful results for breast cancer detection.

Table 2.1. The summary of the previous studiesthat uses characteristic features of images for BC diagnosis

Research	**Dataset**	**Classification method**
Christobel and Sivaprakasamm (2011)	WBC	C4.5, NB, SVM and KNN
Fakoor et al. (2013)	Gene expression	PCA and DL
Ramadevi et al. (2015)	BC,WPBC,WBC,WDBC	PCA and KNN, SVM, C4.5, LR and RF
Maysanjaya et al. (2018)	WDBC	Wrapper and NB
Khourdifi and Bahaj(2018)	WDBC	SVM, RF, NB, and KNN
Keleş(2019)	Antenna	RF, JRip, Bagging, NB, BayesNet
Maliha et al. (2019)	Their original dataset	KNN, J48, NB
Bayrak et al. (2019)	WBC	SVM, ANN
Israni(2019)	WDBC	PCA and Tree, SVM, KNN, SGD, RF, NN, NB, and AdaBoost
Islam et al. (2020)	WDBC	SVM, RF, ANNs, KNN, and LR

2.2. Medical Image Classification Approaches

In this part, the recent studies on BC classification using mammography images are reviewed.

Fuzzy Logic Based Approaches

Some studies use fuzzy logic (FL) -based techniques to diagnose BC. Some of these studies (Baharuddin et al., 2013; Nilashi et al., 2017; Sizilio et al., 2012; Balanica et al., 2011; Sahria, 2019) use only FL or combine it with traditional ML models, while others (Arora and Tagra, 2012; Keleş et al., 2011; Nilashi et al., 2017) use neuro-fuzzy systems for BC. For example, in Baharuddin et al.(2013), the Breast Imaging Reporting and Data Systems(BI-RADS) category of mammograms with mass size and calcification as input data into the fuzzy system was determined by a Mamdani fuzzy expert system. In (Sizilio et al., 2012), the characteristics of breast mass smears obtained by fine needle aspiration (FNA), such as area, circumference, etc., are used as input data in their system. WDBC were used as data set and trapezoidal membership functions (MF) and defuzzification methods with centroid were applied. In another study (Nilashi et al., 2017), the authors developed a knowledge-based BC classifier using classification and regression trees (CART) to generate fuzzy rules. They used a combination of fuzzy rules, PCA, CART and expectation maximization (EM) and tested the knowledge-based system using WDBC and mammography mass datasets. They also showed that the use of clustering techniques, noise reduction, and fuzzy approaches lead to good prediction accuracy in BC diagnosis.

Other studies attempt to calculate the BC risk valuewith a FL approach. In (Balanica et al., 2011), BC risk values are calculated using a selected set of fuzzy rules that take into account the patient's age and tumor characteristics. The proposed model was tested on datasets of 60 patients, each of which included segmentation of tumor results from mammography images, patient's age, and pathological and clinical truth finding by experts. The results from FL appeared to be consistent with the clinical truth when the BC risk score calculated via the

system was compared to the clinical truth. In (Sahria, 2019), BC symptoms such as breast mass, size, shape, and skin lesions were analyzed. These symptoms were then used as data input for a FIS to compute a risk score.

Some other researches employes a combination of DNNs and FL using mammography images to diagnose BC. One such research is (Arora and Tagra, 2012) in which the success of ANFIS was compared with a fuzzy system for detecting BC on a Wisconsin dataset. Several FIS and MFare used by the researchers to validate the results of ANFIS and those of classification. In (Keleş et al., 2011), the authors developed a rule-based neuro-fuzzy model for diagnosing BC on mammograms from the UCI machine learning repository. Their model includes an input, an output, and a hidden layer containing the fuzzy rules.Usingthese rules, they designed an expert system for diagnosis of breast cancer (ex-DBC) inference engine. In another study (Keleş and Keleş, 2013), artificial intelligence methods were used to obtain fuzzy BC classification rules. The researchers used NEFCLASS, a neuro-fuzzy classifier tool, to generate very good fuzzy rules and tested the tool on the mammography mass dataset. Table 2.2 summarizes the previous studies on the fuzzy approach.

In this thesis, we developed a hybrid neuro+fuzzy rule-based system to calculate and classify BC risk usingnewly collected mammography data. Our hybrid model differs from previous studiesby using DL (CNN and ResNet), while other neuro-fuzzy systems (Arora and Tagra, 2012; Keleş et al., 2011; Keleş and Keleş, 2013) test the performane of structures from conventional neural systems. In contrast to the studies of (Baharuddin et al., 2013; Keleş et al., 2011; Keleş and Keleş, 2013; Sizilio et al., 2012; Balanica et al., 2011), in which BC symptoms and patient age are used as input, our proposed model uses mammography images. The system proposed in (Negnevitsky, 2004) is similar to our hybrid system in that (Negnevitsky, 2004) a neural network with backpropagation is applied to the images from SPECT, in our work a hybrid system of neural networks and fuzzy rules, called risk-based hybrid model, is applied to mammograms. As in

(Negnevitsky, 2004), a fuzzy approach is used to increase neural networkperformance. In contrast to (Negnevitsky, 2004), our system computes BC risk using a fuzzy rule-based system.

Table 2.2. The summary of the previous studies for fuzzy approaches.

Study	**Method**	**Task**
Negnevitsky (2004)	Neuro-fuzzy	Myocardial detection
Balanica et al. (2011)	Fuzzy system	BC risk
Keleş et al. (2011)	Neuro-fuzzy and ex-DBC	BC detection
Arora and Tagra(2012)	ANFIS and Fuzzy	BC detection
Sizilio et al.(2012)	Fuzzy system	BC detection
Keleş and Keleş(2013)	Neuro-fuzzy (NEFCLASS)	BC detection
Baharuddin et al.(2013)	Mamdani-Fuzzy expert system	BC detection
Nilashi et al.(2017)	Expectation Maximization, PCA, CART, and Fuzzy Rule-based	BC detection
Sahria(2019)	Fuzzy system	BC risk

CNN-based Approaches:

Convolutional Neural Networks (CNNs) have been used for many image classification problems, especially in medical image analysis. They need huge data and time to train. Nevertheless, occasionally the dataset may be restricted and not sufficient to train a CNN from scratch. In such cases, usage of a pre-trained CNN, which has been trained on a large dataset, may be helpful.

Generally, medical image datasets are limited in size. So, applying a deep learning based approach to solve problems involving medical images is a challenging task. Training CNN with inadequate amounts of training data leads to overfitting, therefore causing incorrect diagnoses.

Transfer learning is the most common techniques to overcome limited training data problem. Transfer learning is a method that stores knowledge gained from one problem solving and applies it to a different task. In other words, it trains a pre-trained model and transfers learned features to a new problem. Convolutional Neural Networks that have been pretrained on ImageNet dataset have different architecture models such as VGG16, Resnet50, InceptionV3, Xception, and

AlexNet. ImageNet dataset has been used to train CNN architecture models. The aim of this image classification challenge is to train a model that can correctly classify an input image into 1000 separate object categories. These image categories represent object classes such as dogs, cats, several household objects, vehicle types, and much more.

In the literature, researchers have used diverse pre-trained models for transfer learning to analyze medical image data.

Spanhol et al. (2016) classified lesion images using Inception-V3 architecture in which the weights are fine-tuned by resuming backpropagation on all layers with a per-layer exponentially decaying learning rate.

Chougrad et al. (2017) applied VGG16 model for feature extraction from mammogram images on mini-MIAS dataset. The extracted features are used to train a new neural network by using the DDSM dataset. The weights in the final layers of the VGG16 model are updated with a mini-MIAS data by backpropagation to detect abnormal regions.

It has been seen from (Chang et al., 2017), the authors proposed a transfer learning technique to detect breast cancer using histopathology images based on Google's InceptionV3 model. In order to achieve higher accuracy, in addition to transfer learning, they applied data augmentation methods to have enough data in training. Their model performed classification with an accuracy of 83 % for benign class and 89 % for malign class.

As it has been proposed in (Meirom et al., 2017), researchers have used InceptionV3 model to classify mammogram images. Additionally, a method called TI Pooling was implemented to increase the model's ability to correctly classify the scans. That method increased the model's accuracy from around 82 % to 87.6 %.

Wu et al. (2018) used ResNet34 pre-trained model to classify breast cancer on DCE-MRI. The key idea of ResNet architecture is residual blocks, in which an identity connection is added parallel to stacked nonlinear layers. This improves gradient flow through the network during backpropagation allowing training very

deep models. Before transfer learning step, they augmented data using random rotations and flipping, and resized images to 224x224 pixels. And also, the final layer of ResNet architecture was adapted to the binary classification problem to classify DCE-MRI data as malignant or benign. The accuracy that has been obtained by ResNet model is around 80%.

Similarly, in (Haarburger et al., 2018), the researchers obtained accuracy around 80% using AlexNet model to identify masses in mammograms. The network is fine-tuned by modifying the last few layers and keeping the initial layers of the pretrained network as is, so as to learn the specific features of images in the new data set. They increased size of data by image rotation and mirroring. The model has been adapted to identify 2 classes instead of 1000 classes by modifying corresponding layers. The results show that data augmentation increases the learning accuracy when using transfer learning. With the data augmentation, model is having accuracy around 80 %.

According to study (Motlagh et al., 2018), the authors compared the performance of different structures of ResNet and Inception to classify breast cancer histopathological images. The migration of inception models is from fully networked to sparsely networked architectures. Nonlinearity capability is added to inception models through 1x1 factorised CNNs, followed by Rectified Linear Unit (ReLU). Shortcuts between shallow and deep networks are always used by ResNet to monitor and modify the training error. The researchers in this study examined several frames such as (V150, V1 101, and V1 152) of ResNet and (V1, V2, V3, and V4) of Inception using breast cancer data. The performance results were 98.4 %, 97.8 %, and 94.1 % for ResNet V1 101, ResNet V1 50, and Resnet V1 152 respectively. Compared to other Inception architectures, Inception V2 achieves higher accuracy of 94.1 %. In conclusion, this experiment shows that the ResNet model is a reasonable and valid method for diagnosing cancer based on certain parameters compared with other conventional models.

As can be seen in study (Hassan et al., 2020), AlexNet and GoogleNet pre-trained models have been used to classify breast cancer masses. The authors compare the results of two models on CBIS-DDSM, NCI, MIAS, and INbreast datasets. According to the experimental results, while the lowest classification accuracy of AlexNet model is 97.89%, the lowest classification accuracy of GoogleNet is 88.24%. Overall results indicate that AlexNet has better performance than GoogleNet.

According to Falconi et al.(2020), the authors classify mammogram images using CBIS-DDSM database. They make a large comparison of transfer learning and fine-tuning techniques with different ConvNet models. They show that VGG model has better score compared to ResNet model in fine-tuning.

As can be seen from the above studies, convolutional neural networks have been widely used to classify breast cancer. This is because CNNs show very good performance in extracting useful patterns from images. On the other hand, in addition to their excellent performance in feature extraction, CNNs also have some weaknesses. They are generally unable to focus on local variations in image patterns, so they lose the global information of the features (Parvaiz et al., 2022).

Transformer Model Approaches:

Recently, transformers have become popular as they focus on local variations in images and improve the performance of neural network models (Y. Liu et al. 2019; Han et al. 2021). In the literature, there are two ways of using transformers. One is to use only pure transformers (Dosovitskiy et al. 2021). The second is the use of a combination of CNN and transformer, such as the bottleneck transformer (Srinivas et al., 2021).

TransMIL was proposed by Shao et al. (2021) for whole-slide image classification. It is based on the pure transformer method and uses three different cancer datasets: Breast cancer (CAMELYON16) (Veeling et al., 2018), Lung cancer (TCGA-NSCLC) (Napel and Plevritis, 2014), and Kidney cancer (TCGA-

R) (National Cancer Institute, 2022). Its performance approaches the results of the current state of the art using these data sets.

Chen et al. (2022) developed a transformer-based method for classifying mammograms with multiple views. They compare the performance of the transformer model with computer-aided diagnosis methods for multiview images and show that the results of the transformer model are better in malignancy classification. Thus, the proposed model may be a better alternative than CNNs for multiview images.

ScoreNet (Stegmueller et al., 2022) is a type of transformer model for classifying histopathological breast cancer using the BRACS dataset (Pati et al. 2022). The ScoreNet results have a fairly high score.

Table 2.3 shows a summary of previous studies that used the Transfer Learning (TL) and Transformer models for BC classification. As shown in the table, both (Chang et al., 2017) and (Motlagh et al., 2018) use histopathological images, while others use mammogram images for BC classification. TL is commonly used for breast cancer classification because the amount of data in medical is small. On the other hand, the Transformer model is the latest model for mammogram classification, as shown in the Table 2.3. There are few studies (Shao et al., 2021; Chen et al., 2022; Stegmueller et al., 2022) using the Transformer model.

Apart from the above studies, similar to (Chougrad et al., 2017; Haarburger et al., 2018; Falconi et al., 2020), we used the TL method with ResNet50, AlexNet, and VGG16 to classify newly collected mammograms because we had a limited amount of training data, and we found that the accuracy in classifying mammogram images of breast cancer was higher compared with other models. However, transfer learning (ResNet50) with a fuzzy logic approach was used to classify newly collected mammogram images, which makes our study different from previous studies. In addition, the performance of Resnet50+fuzzy was compared with that of CNN+fuzzy in classifying newly acquired mammograms.

Since the focus was on local variations in the images, we also analyzed the mammogram images using a transformer model similar to (Chen et al., 2022). In contrast to previous studies, we compared the performance of the transformer model with that of pre-trained models (ResNet50, AlexNet, and VGG16). We also tested mammograms using a hybrid of a TL and a Transformer model. For this purpose, we use use VGG16 to obtain inputs for the transformer model. To our knowledge, there is no study that uses VGG16+Transformer to classify mammograms.

Table 2.3. The summary of the previous studies usingCNN-basedmethods and Transformers.

Study	**Method**	**Classification Task**
Spanhol et al. (2016)	Inception-V3	Mammogram
Chougrad et al.(2017)	VGG16	Mammogram
Chang et al.(2017)	Google's InceptionV3	Histopathology
Meirom et al.(2017)	InceptionV3 and TI Pooling	Mammogram
Wu et al.(2018)	ResNet34	MR
Haarburger et al.(2018)	AlexNet	Mammogram
Motlagh et al.(2018)	ResNet and Inception	Histopathology
Hassan et al.(2020)	AlexNet and GoogleNet	Mammogram
Falconi et al.(2020)	VGG and ResNet	Mammogram
Shao et al.(2021)	TransMIL	Mammogram
Chen et al.(2022)	Tansformer	Mammogram
Stegmueller et al.(2022)	ScoreNet (Transformer based)	Histopathology

2.3. Radiology Reports Classification Approaches

Several text classification and NLP techniques have been applied to the medical domain, e.g., sentiment analysis, radiology report classification, and medical article classification. Most of these tasks use English texts, so we can conclude that there are a limited number of studies on radiology report analysis in a language other than English. In this part, the literature was searched for text classification methods and studies in other diseases in the radiology report are as follows:

In the literature, there are a few studies in Turkish language including Arifoğlu et al. (2014), Parlak and Uysal (2020),Çelıkten and Bulut (2021). The authors in (Arifoğlu et al., 2014) construct an International Classification of Diseases (ICD-10-AM) coding system for patient records in Turkish hospitals. ICD-10-AM codes show the symptom or classification of a disease and used to find the classification of a disease. Automatic assignment of ICD codes to patient records in Turkish requires expert knowledge and is also error-prone, as human annotators must consider thousands of possible codes when assigning the correct ICD label to a document. They employ the bag-of-words approach using a Lucene search engine and the Borda counting method, and show that their results are successful in automatically assigning ICD-10-AM codes to clinical findings. Çelikten and Bulut (2021) examine the problem of classifying Turkish medical texts using BERT models. In this study, article summaries belonging to 10 different disease groups and 2 BERT-based models are used for the classification problem of medical texts. These models attempt to assign the article summaries to the appropriate disease category. They achieve an F score of 0.82 and 0.93 for the multilingual BERT and BERTurk, respectively. The results show that the BERTurk model is more successful than the other compared models for Turkish medical text classification. The authors Parlak and Uysal (2020) examine a comprehensive comparison of the classification of Turkish and English counterparts of the same abstracts published in Turkish medical journals. They use unigram, bigram, and hybrid (the combination of unigram and bigram) methods to extract features. The best results are obtained from the combination of unigram features, distinguishing feature selector (DFS) and multinomial naïve Bayes (MNB) classifier for both data sets.

The studies in the literature for languages other than Turkish language are as follows:

In Shin et al. (2017), the researchers propose a novel neural attention mechanism where convolutional neural networks (CNN) with attention analysis are

used to classify radiological head computed tomography reports. Their experiments show that CNN attention models perform better than non-neural models.

Castro et al. (2017) focus on a rule-based BI-RADS Observation Kit (BROK) algorithm and Bayesian networks to classify mammography reports. BROK uses regular-expression based string matching to obtain the reports' BI-RADS category and the laterality of the breast to which it is assigned, when possible. BI-RADS (Breast Imaging-Reporting and Data System) (Niknejad, 2022) is a risk assessment and quality assurance tool developed by American College of Radiology (Bell, 2020) that provides a widely accepted lexicon and reporting schema for imaging of the breast. They first evaluate BROK algorithm to extract BI-RADS categories data from mammography reports, and then use this data as input to a Bayesian network to categorize radiology reports as malignant and benign according to the probability provided by the Bayesian network.

A recent study by Saib et al. (2020) presents a hierarchical CNN classification method applied to the prediction of ICD-O codes for pathological breast cancer reports. The performance of hierarchical CNN is compared with the performance of multiclass CNN. The results show that a multiclass CNN has F1 macro and F1 micro scores of 0.717 and 0.738, respectively, compared to a hierarchical CNN with F1 macro and F1 micro scores of 0.722 and 0.748.

Nguyen et al. (2020) develop a hybrid system of a language model and a BI-RADS -score classifier to automatically summarize Dutch radiology reports of breast cancer. BI-RADS apply to mammography, ultrasound, and MRI. Similar to Arifoğlu et al. (2014), they use a bag-of-words method and obtain TF-IDF values for word features. In short, they create a TF-IDF matrix from radiological reports and use it as input to machine learning algorithms to determine the score of BI-RADS. The SVM algorithm proves to be the most successful in predicting the BI-RADS score with an accuracy of 83.3% compared to the other algorithms.

Magna et al. (2020) focuses on breast cancer diagnosis using free medical text. They apply word2vec and BERT methods for converting text data into word

vectors. Moreover, the results obtained using word2vec and BERT are compared with the technique TF-IDF. Then, these word vectors are classified using machine learning and deep learning models. SVM, Decision Tree (DT), NB, KNN, RF are used as machine learning algorithms and Multiplicative Long Short Term Memory (LSTM), Dense, Bidirectional LSTM are used as Deep Learning algorithms. The experimental results show that the combination of word2vec and Bidirectional LSTM with a macro F1 score of 98% performs better than all others in classifying breast cancer in the Spanish dataset.

Another study on radiology reports is prepared by Casey et al. (2021). The researchers present a systematic review of publications using NLP on radiology reports (2015-2019) following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) (Moher et al., 2015).

A recent study by Boroumandzadeh and Parvinnia (Boroumandzadeh andParvinnia, 2021) proposes an approach consisting of five main parts, namely text report processing, feature extraction, feature engineering, prediction, and evaluation for automatic extraction of BI-RADS classifications from text reports. In the first part, the information is extracted from medical reports using a mammography dictionary. In the second step, the combination of Word2Vec and TF-IDF is used to generate a feature vector for each medical report based on the extracted information. In the third step, the patient records are added to the feature vectors generated by Word2Vec and TF-IDF, called HIS. In the fourth step, the resulting feature vectors are classified using SVM, NB, Extreme Gradient Boosting (XGBoost) and Multilevel Fuzzy Min-Max Neural Network (MLF) classifiers. In the last step, the results of with-HIS are compared with the results of without-HIS in terms of accuracy. The experimental results show that MLF is more accurate than other algorithms. The accuracy of MLF for HIS is 89%, while the accuracy of MLF without HIS is 85%.

Gupta et al. (2021) propose a method to classify gene mutations based on the text description of these genetic mutations for cancer prediction using Natural

Language Processing (NLP) techniques. In this study, the texts are transformed into matrix token counts using the Word2Vec, TF-IDF Vectorizer, and CountVectorizer transformation models, and this sparse matrix is then classified using the LR, RF, XGBoost, and Recurrent Neural Network (RNN) models. The experimental results show that RNN outperforms all other algorithms with 70% accuracy.

Another recent study is Suárez-Paniagua et al. (2021) in which the authors use radiology reports written in Spanish and analyze them using a hybrid entity recognition system and BERT. They demonstrate the performance of combining the hybrid system with multiple BERT and Google Healthcare Natural Language API over the documents translated into English with cross-language word matching, and this technique achieves the best recall in the task.

Grancharova and Dalianis (2021) investigate the fine-tuning of a Swedish BERT model and a multilingual BERT model for NERC on a Swedish corpus of electronic patient records. The experimental results of their study show that the Swedish model is very successful with a recall of 0.9220 and a precision of 0.9226.

The authors in (Yogarajan et al., 2021) use transformers to improve the prediction of medical codes using Medical Information Mart for Intensive Care (MIMIC-III) and Electronic Intensive Care Unit (eICU) datasets. They show that the performance of the new transformer, Longformer, is better than standart transformers for long medical documents.

In (Nakamura et al., 2021), the authors investigate automatic detection of radiology reports with actionable findings using BERT. They use radiology reports in Japanese for computed tomography examinations (CT) performed at the researchers' hospital. They compare the results of BERT with the baseline LSTM and gradient boosting decision tree methods. The results show that BERT performs better than the other baseline models in discriminating between actionable and non-actionable radiology reports.

Jaiswal et al. (2021) present a new method RadBERT- CL, a pre-trained model that uses contrastive learning to improve the performance of SOTA transformers (BERT, BlueBERT, etc.) in capturing critical medical and factual nuances of radiology reports. They use the dataset MIMIC-CXR (Johnson et al., 2019), which consists of chest radiographs with corresponding de-identified radiology reports. The results show that the RadBERT- CL method, i.e., pre-training transformers through contrastive learning, provides significant improvement in fine-tuning tasks compared to SOTA transformers BERT \BlueBERT. They show that RadBERT- CL is very successful and outperforms BERT \BlueBERT on classification problems.

In (Yan et al., 2022), the authors use transform-based language models for radiology reports obtained from U.S. Department of Veterans Affairs health systems. They create six variants of RadBERT pretraining reports on Clinical-BERT, BioMed-RoBERTa, and BERT -base. Each variant was fine-tuned for three tasks, abnormal classification, report coding, and report summarization, and compared to baseline models such as BERT -base, BioBERT, Clinical- BERT, BlueBERT, and BioMed-RoBERTa. When the performance of the baseline models and the RadBERT variants for abnormal classification is analyzed, the results of the RadBERT variants show a higher F1 score and accuracy rate even when the RadBERT variants are trained with 10% or less training data. Similarly, the RadBERT variants also show better performance than the baseline models on other tasks.

Table 2.3 shows a summary of previous research with medical text data. As shown in the table, (Castro et al., 2017) and (Parlak and Uysal, 2020) use BROK and Unigram and Bigram, respectively, to extract features. These studies are different from other studies that use different feature extraction methods, such as TF-IDF, Countvec, and Word2Vec. Nguyen et al.(2020), Magna et al. (2020), Gupta et al. (2021) and Boroumandzadeh and Parvinnia (2021) use vectorizers (TF-IDF, CountVec, Word2Vec) that convert medical texts into numerical vectors

and use these vectors as input to the ML algorithms. Çelıkten and Bulut (2021) and Grancharova and Dalianis (2021) use BERTMultilingual and BERT native language variants to classify medical texts in their native language. Yogarajan et al. (2021) differs from previous studies by using transformers. Both Jaiswal et al. (2021) and Yan et al. (2022) use the RadBERT base method for medical data. Jaiswal et al. (2021) combines RadBERT with contrastive learning, while Yan et al. (2022) uses variants of RadBERTs.

Table 2.4. The summary of the previous studies for radiology reports.

Study	**Language**	**Classification problem**	**Feature extraction**	**Classifier**
Arifoğlu et al. (2014)	Turkish	ICD codes	Bag of words approach	Lucene search engine and the Borda counting method
Shin et al. (2017)	English	CT head reports	-	CNN attention and Non-neural models
Castro et al. (2017)	English	Mammography reports	BROK	Bayesian Network
Saib et al. (2020)	English	ICD-O codes		CNN
Nguyen et al. (2020)	Dutch	BC radiology reports	TF-IDF	SVM and other ML algorithms
Magna et al. (2020)	Spanish	BC medical text	TF-IDF	BERT, SVM, DT, NB, KNN, RF LSTM
Parlak and Uysal (2020)	Turkish and English	Medical abstracts	Unigram, Bigram	DFS and MNB
Casey et al. (2021)	English	Radiology reports	-	PRISMA
Boroumandzadeh and Parvinnia (2021)	English	Mammography reports	Word2Vec and TF-IDF	SVM, NB, XGBoost, MLF
Gupta et al. (2021)	English	Gene mutation	Word2Vec, TF-IDF Vectorizer, and CountVectorizer	SVM, LR, RF, XGBoost, and RNN
Suárez-Paniagua et al. (2021)	Spanish	Ultrasonography reports	-	Hybrid entity recognition system, BERT

Çelıkten and Bulut (2021)	Turkish	Medical texts	-	BERTMultilingual and BERTTurk
Grancharova and Dalianis (2021)	Swedish	Electronic patient records	-	BERTSwedish and BERTMultilingual
Yogarajan et al. (2021)	English	Medical texts	-	Transformers
Nakamura et al. (2021)	Japanese	CT reports	-	BERT, LSTM, DT
Jaiswal et al. (2021)	English	Chest radiograph reports	-	RadBERT-CL, BERT, BlueBERT
Yan et al. (2022)	English	Radiology reports	-	Variants ofRadBERT

Classification of medical texts is an important research problem for Turkish because it has a compelling morphological structure and many specified words occur in medical texts. As shown in Table 2.4, there are very few studies in the medical field that deal with text classification in Turkish. In this thesis, similar to previous studies, we use TF-IDF, to convert words into vectors and classify newly created Turkish radiology reports using ML algorithms. Similar to (Çelıkten and Bulut, 2021), we also use BERTTurk and BERTMultilingual to classify newly created reports. In addition, BERTClinical is used to investigate the impact of the pre-trained model with clinical texts on our newly created data. We also use Transformers as well as ML algorithms and BERTs to classify newly created radiology reports. To our knowledge, this is the first study to use Transformers with BERTMultilingual to classify newly created Turkish radiology reports.

3. MATERIALS AND METHODS

In this section of the thesis, the materials and methods used in this thesis study are explained in detail.

3.1. Materials

3.1.1. Collection of the Datasets

One of the main objectivefor this study is to collect datasets from radiology reports, surveys, and mammographic images from each patient. For this reason, with the decision of the Ethics Committee for Non-Interventional Clinical Research of Çukurova University, a survey was created to collect data from patients' clinical reports, including medical history, mammogram images, and radiology reports. Fig. 3.2 shows the workflow of data collection.

At the beginning of data collection in the hospital, we first create surveys with the help of a general surgeon. First, surveys are applied to patients who came to the Department of General Surgery with breast complaints. A total of 169 patient surveys were completed. Second, using the patient numbers provided in the surveys, we collect the corresponding mammogram images in digital imaging and communications in medicine (DICOM) format from the hospital's image archiving system (PACS) and the corresponding radiology reports.

Because of the lack of image data, we were able to obtain mammogram images and the corresponding radiographic reports from only 62 patients. However, for ethical reasons, we cannot share these collected data sets with the literature. We could not include the image data of the other patients in the surveys because they may have had their mammograms obtained elsewhere. 42 malignant and 20 benign mammograms of 62 patients were acquired with the Fujifilm Amulet Innovality B-115 mammography machine at the radiology department of Çukurova University Faculty of Medicine.

Subsequently, the Information Technology Department (IT) removed all patient identification data such as patient name, address, date of birth, etc. from both the images and the associated radiology reports. Labelling of images, exams, and radiology reports are performed by medical experts. The MicroDICOM image viewer is used for Dicom images. The image size of mammograms is 4728x5928 pixels. Powerful computers are required to work with these larger Dicom images. To solve this problem, we convert the images from DICOM to PNG format and resize them.

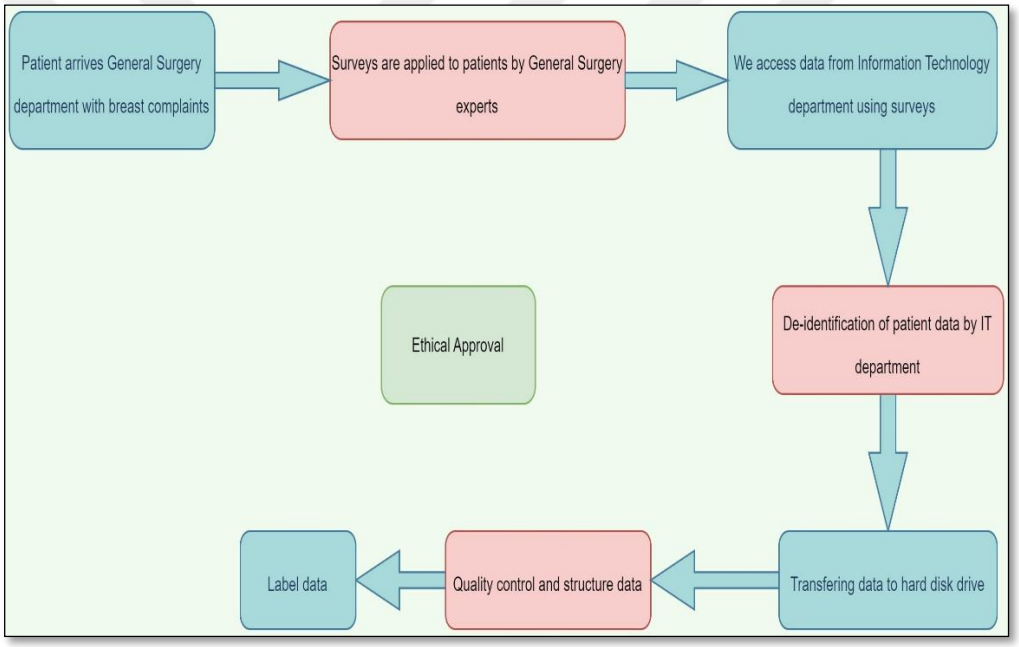


Figure3.1. Theworkflow of dataset (image, text, and survey) collection.

3.1.2. Characteristics of the Hospital Dataset

Our dataset consists of three different types of data: Text, Image, and Surveys. Table 3.1 shows the number of samples in the hospital dataset.

Table3.1. The number of samples in the hospital dataset.

Class	Survey	Image	Text
Malignant	126	42	42
Benign	43	20	20
Total	169	62	62

3.1.2.1. Survey dataset

All survey questions were prepared by an expert from the Department of General Surgery at our university hospital. The survey was developed using Google Forms, which can be used free of charge and has been used in patients with breast complaints. With Google Forms, we can insert any type of question into the survey, organise it by adding headings, or divide it into severalpages. The survey consists of three sections: Medical History, Physical Examination, and Radiological Imaging. Each section contains different question types, including text boxes, multiple-choice questions, drop-down lists, and checkbox questions. By applying the survey to each patient, we collect some information about the patients' Medical History, Physical Examination, and Radiological Imaging. In this way, a total of 169 surveys were completed. Then, we downloaded the surveys completed by the experts from the system into an Excel file. Finally, we transferred the responses to the surveys by assigning a numerical value to each response to the questions that did not contain numerical values.

Table 3.2 shows the survey questions, answer data types, and from which part of the survey that question comes. As shown in Table 3.2, our newly created survey consists of 56 questions, and each questionhas response with a binary, real, or multiple data types.

Table 3.2. Surveyquestions, response data types, and obtained from which part of the survey.

Survey questions	**Type**	**Obtained from**
Gender	Binary	Patient History
Age	Real	Patient History
Weight (kg)	Real	Patient History
Height (cm)	Real	Patient History
Breast density	Binary	Patient History
Hormon Replacement Therapy (HRT)	Binary	Patient History
Oral Contraceptive Use (OCS)	Binary	Patient History
Menarche age	Binary	Patient History
Number of births	Binary	Patient History
Infertility	Binary	Patient History
First birth age	Binary	Patient History
Patient breast cancer story	Binary	Patient History
Family breast cancer story	Multiple response	Patient History
Genetic mutation	Binary	Patient History
Alcohol usage	Binary	Patient History
Smoking	Multiple response	Patient History
Employee after midnight	Binary	Patient History
Menopause age	Binary	Patient History
Exposue to ionizing radiation	Binary	Patient History
Breastfeeding time	Binary	Patient History
Physical activity	Binary	Patient History
Diet	Binary	Patient History
Previous biopsy	Binary	Patient History
Use of aspirin	Binary	Patient History
Orange image	Binary	Pysical Examination
Nipple retraction	Binary	Pysical Examination
Ulceration of the skin	Binary	Pysical Examination
Satellite nodule	Binary	Pysical Examination
Oophorectomy	Binary	Pysical Examination
Concomitant axillary lymphadenopathy	Binary	Pysical Examination
Hard mass	Binary	Pysical Examination
Skin retraction	Binary	Pysical Examination
Irregularly limited mass	Binary	Pysical Examination
İmmobile mass	Binary	Pysical Examination
Painless mass	Binary	Pysical Examination
Erythematous squamous lesion on the nipple	Binary	Pysical Examination
Spiculation	Binary	Radiological Imaging
Lesions longer than width	Binary	Radiological Imaging
Irregular boundary faint contour	Binary	Radiological Imaging
Posterior shadowing	Binary	Radiological Imaging
Hypoechoic lesions	Binary	Radiological Imaging
Lesions with calcification	Binary	Radiological Imaging
Ductus enlargement	Binary	Radiological Imaging
Microbulation	Binary	Radiological Imaging

Branching pattern	Binary	Radiological Imaging
Increased blood supply within the lesion	Binary	Radiological Imaging
Angular edge	Binary	Radiological Imaging
Microcalcification	Binary	Radiological Imaging
Asymmetrical density	Binary	Radiological Imaging
Opacity	Binary	Radiological Imaging
Spicular mass	Binary	Radiological Imaging
Radial stretching	Binary	Radiological Imaging
BIRADS	Multiple response	Radiological Imaging
Type3 contrasting	Binary	Radiological Imaging
Type2 contrasting	Binary	Radiological Imaging
Type1 contrasting	Binary	Radiological Imaging

Answer of each question form an attribute of the dataset. Some statistics for each attribute of the survey are shown in Table 3.3. Some statistics such as minimum, maximum, mean, standard deviation, median, and range of the attributes listed in Table 3.3 were calculated using Equations 3.1, 3.2, 3.3, and 3.4.

$$Range = Highest\ vale - Lowest\ value \tag{3.1}$$

$$Mean = \frac{Sum\ of\ values}{Number\ of\ values} \tag{3.2}$$

$$Median = (\frac{N+1}{2})^{th}\ term \tag{3.3}$$

$$Std.deviation = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\mathrm{N}}} \tag{3.4}$$

Where, N is total number of patients and μ is mean.

Table 3.3.Some statistics on each attribute in the survey dataset.

Surveyquestions (attribute)	**Mean**	**Median**	**Std. Deviation**	**Range**	**Minimum**	**Maximum**
Age	49,947	49	12,661	57	18	75
Weight	72,231	70	14,514	80	40	120
Height	160,14	160	9,9435	118	62	180
Breast density	0,8107	1	0,7152	2	0	2
Hormone Replacement Therapy (HRT)	1,5385	2	0,5564	2	0	2
Oral Contraceptive Use (OCS)	0,716	1	0,4902	2	0	2
Menarche age	1,0118	1	0,8863	3	0	3
Number of births	0,9882	1	0,5231	2	0	2
Infertility	0,9941	1	0,2988	2	0	2
First birth age	1,7456	2	0,802	3	0	3
Patient breast cancer story	0,8047	1	0,4665	2	0	2
Family breast cancer story	0,5621	0	1,3309	8	0	8
Genetic mutation	1,3905	1	0,5681	2	0	2
Alcohol usage	0,9527	1	0,213	1	0	1
Smoking	1,0769	0	2,5542	13	0	13
Employee after midnight	0,9467	1	0,3136	2	0	2
Menopause age	1,8698	2	1,178	3	0	3
Exposue to ionizing radiation	0,9467	1	0,294	2	0	2
Breastfeeding time	0,7633	1	0,7093	2	0	2
Physical activity	0,4615	0	0,5	1	0	1
Diet	0,7633	1	0,4401	2	0	2
Previous biopsy	0,7633	1	0,6101	2	0	2
Use of aspirin	0,8462	1	0,378	2	0	2
Orange image	0,9763	1	0,3077	2	0	2
Nipple retraction	0,8757	1	0,4391	2	0	2
Ulceration of the skin	0,9822	1	0,2983	2	0	2
Satellite nodule	0,9408	1	0,4843	2	0	2
Oophorectomy	1,8521	2	0,4031	3	0	3
Gender	1	1	0	0	1	1
Concomitant axillary lymphadenopathy	0,645	1	0,5601	2	0	2
Hard mass	0,3609	0	0,5824	2	0	2
Skin retraction	0,8698	1	0,4702	2	0	2

Irregularly limited mass	0,5444	0	0,6264	2	0	2
İmmobile mass	0,5325	0	0,598	2	0	2
Painless mass	0,4852	0	0,5785	2	0	2
Erythematous squamous lesion on the nipple	1,0414	1	0,3154	2	0	2
Spiculation	0,8817	1	0,6885	2	0	2
Lesions longer than width	0,929	1	0,6777	2	0	2
Irregular boundary faint contour	0,7101	1	0,7432	2	0	2
Posterior shadowing	0,9586	1	0,6579	2	0	2
Hypoechoic lesions	0,8047	1	0,7889	2	0	2
Lesions with calcification	0,8698	1	0,6863	2	0	2
Ductus enlargement	1,0888	1	0,5652	2	0	2
Microbulation	1,0828	1	0,6116	2	0	2
Branching pattern	1,1479	1	0,5417	2	0	2
Increased blood supply within the lesion	1,1598	1	0,5809	2	0	2
Angüular edge	1,2189	1	0,5284	2	0	2
Microcalcification	1,0237	1	0,8088	2	0	2
Asymmetrical density	0,8994	1	0,8356	2	0	2
Opacity	1,0355	1	0,8514	2	0	2
Spicular mass	1,0592	1	0,7921	2	0	2
Radial stretching	1,2663	1	0,6682	2	0	2
BIRADS	7,6805	8	2,1968	10	0	10
Type3 contrasting	1,426	2	0,8067	2	0	2
Type2 contrasting	1,5976	2	0,7098	2	0	2
Type1 contrasting	1,6686	2	0,5848	2	0	2

Fig. 3.2, Fig. 3.3, and Fig. 3.4 show the frequency graph of numerical attributes age, weight, and height.

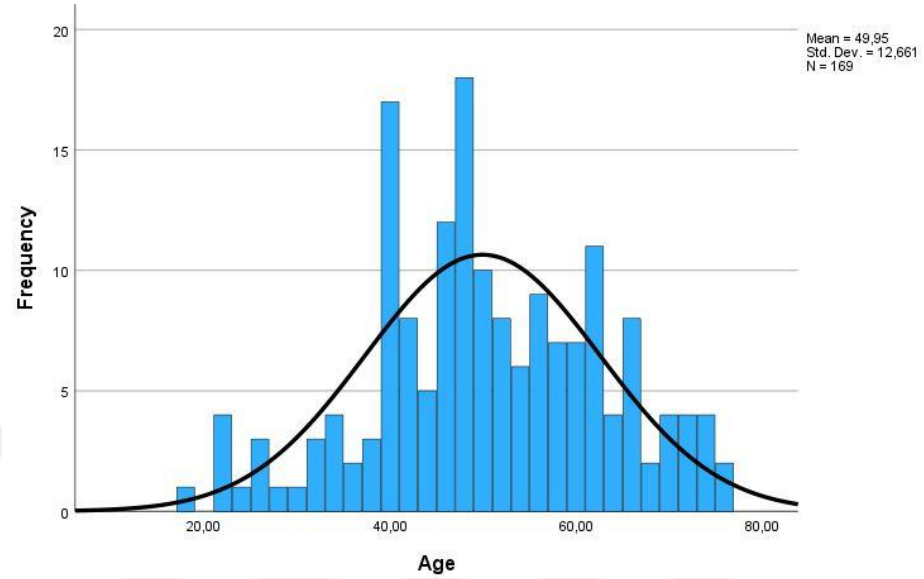


Figure 3.2. The frequency of age.

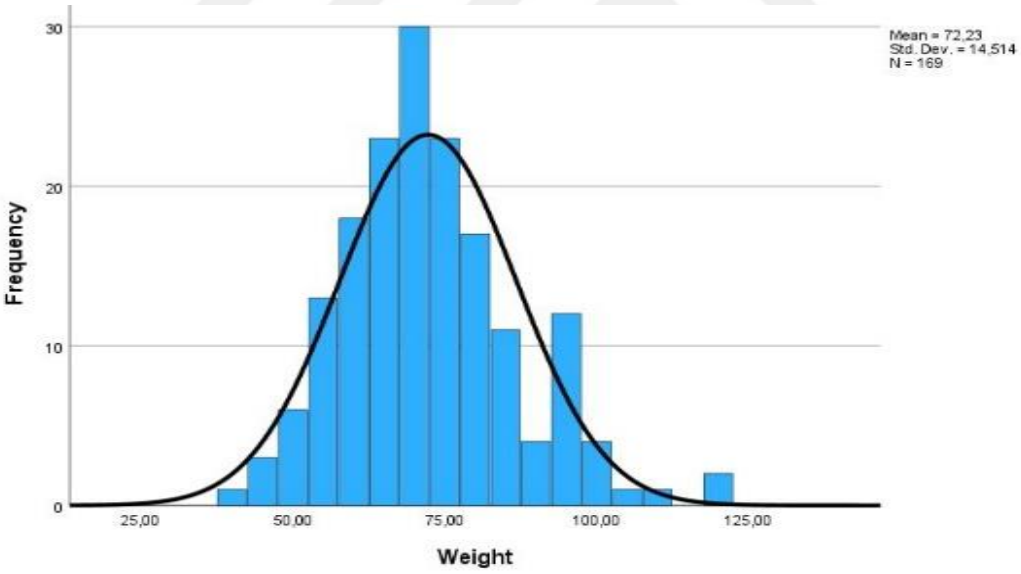


Figure 3.3. Thefrequency of weight.

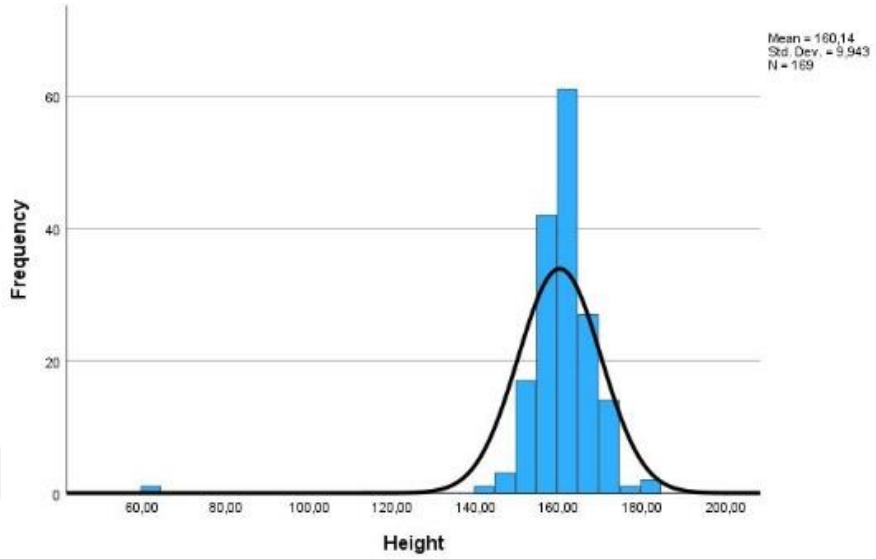


Figure 3.4. Thefrequency of height.

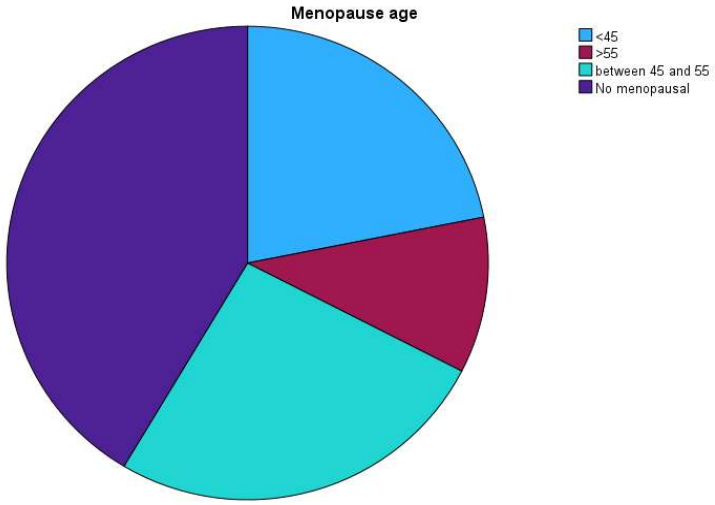


Figure 3.5. Thechart of menopause age.

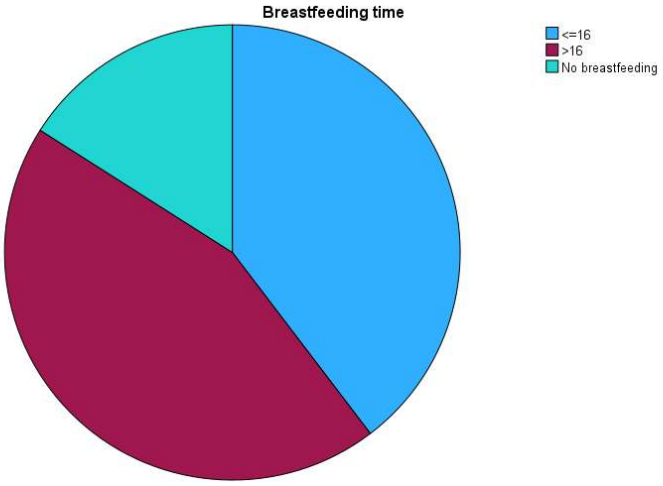


Figure 3.6. Thechart of breastfeeding time.

We also show pie charts of only two attributes to save space. Fig. 3.5and Fig. 3.6show the pie chart of menopause age and breastfeeding time.

3.1.2.2. Mammography image dataset

The images of the dataset were collected from patients in the Department of General Surgery, Faculty of Medicine, Cukurova University. The collected mammography dataset contains 42 malignant and 20 benign mammograms. Fig. 3.7 and 3.8 show the images of the malignant and benign mammograms, respectively. Each mammogram image was acquired using the Fujifilm Amulet Innovality B-115 mammography device. These mammograms contained some patient information. This information was anonymized by the IT department to protect patient privacy. The mammogram images are in DICOM format and have a size of 4728x5928 pixels. The mammogram images were converted to PNG format and resized.

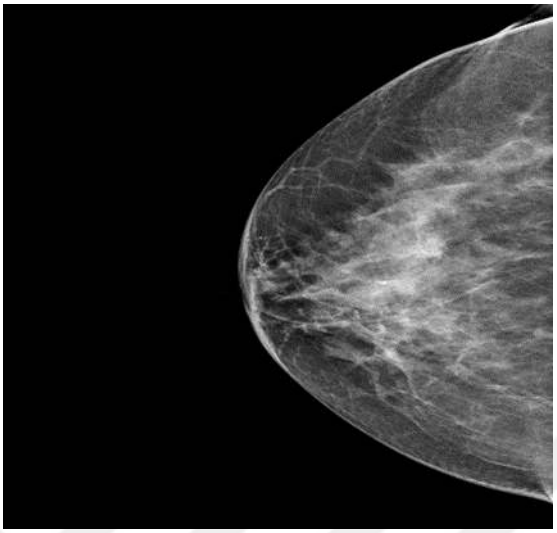


Figure 3.7. The malignant type mammography image.

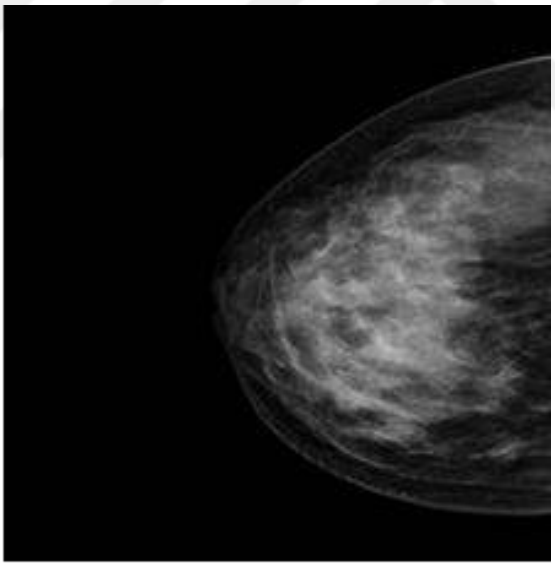


Figure 3.8.The benign type mammography image.

In this work, we also used the BCDR-D02 (Lopez et al., 2012) and mini-MIAS (Suckling, 1994) datasets for some experiments. In BCDR-D02, the number of benign mammograms is 397, while the number of malignant mammograms is 42. In mini-MIAS, there are more than 20 mammograms with multiple

abnormalities, but we only use calcification type mammograms, i.e., 10 malignant and 10 benign mammograms.

3.1.2.3. Radiology report dataset

Our dataset for radiology reports consists of 20 reports labeled as benign and 42 reports labeled as malignant. Table 3.4 shows the number of tokens by each class. Number of tokens in the report with the shortest length in the benign class is 11, in the report with the longest length is 196, and the average of the number of tokens for all reports in the benign class is 106. The number of tokens of the longest report and the shortest report for the malignant type is 311 and 24, respectively. The average of the number of tokens for all reports of the malignant typeis 133. The number of unique tokens ofthe malignant typereportsand benign type reports is 977 and 591, respectively. The total number of unique tokens from all reports is 1167.As can be seen in Table 3.4, the number of tokens in the malignant class is higher than the number of tokens in the benign class for minimum, maximum length reports. This is because more detailed information is given in the malignant reports, such as the type, size, and location of the tumor.

An example of a radiological report from the dataset is shown in Fig.3.9. To understand the characteristic of the dataset, we extracted the 35 most common words from benign and malignant reports, respectively. The clouds of the most common words for benign and malignant reports are given in Fig.3.10and 3.11. It is clearly seen that some words such as “meme" (in English, “breast"), “kitle" (in English, “mass") are common in benign and malignant reports.

Table 3.4. Some statistics for radiology report dataset.

Class label	**# of tokens**			
	min report length	**max report length**	**Average report length**	**Total # of unique tokens**
Malignant	24	311	133	977
Benign	11	196	106	591

Sol meme saat 3 düzeyinde 2x2 cm boyutunda spiküle malign tümöral kitle mevcut olup sol aksilada patolojik boyutta lenf nodu mevcut değildir.
Yine solda saat 8 düzeyinde 5x9 mm boyutunda santrali ekojen periferinde milimetrik kistik alanlar içeren iyi sınırlı lezyon izlenmekte olup sağ meme ve sağ aksilla doğaldır.

SONUÇ: Solda 2x2 cm boyutunda malign tümöral kitle. Sol saat 8 de 5x9 mm boyutunda benign kriterler gösteren lezyon (tip 3 kompleks kist ?) .
BIRADS 6.
Sol saat 8 deki lezyonun takip ve kontrolü önerilir.

Figure 3.9. A Radiology report.

Figure3.10. The most common words obtained from malignant type radiology reports.

Figure 3.11. The most common words obtained from benign type radiology reports.

3.2. Methods

In this section, the methods used in thesis as well as the proposed models of a hybrid approach to diagnose breast cancer and reduce unnecessary breast biopsies using machine learning (SVM, C4.5, Random Forest, etc.), fuzzy approach, and deep learning (transfer learning, BERTs, and transformers) are detailed.

3.2.1. Preprocessing Methods

Since each type of dataset (survey, mammogram, and radiology reports) is too small to train the DL model, we apply K-Fold cross validation (Han et al., 2022) in which we divide each type of data into k distinct subsets, then we use $k-1$ subsets to train our data and leave the remaining subset as test data. The k value is usually given as “5” or “10” in the literature. For our experiments, we choose k to be 5 because our data set is small, and DL methods take too much time.

Before learning a classification model from the survey dataset, the mammography dataset, and the radiology report dataset, we applied some

preprocessing methods to each data type to increasethe success of the classifiers.In addition to augmentation and balancing, feature scaling and selection steps are applied to the survey dataset.

Fig. 3.12, Fig. 3.13, and Fig. 3.14 show the summary of the preprocessing methods for radiology reports, surveys, and mammograms, respectively.

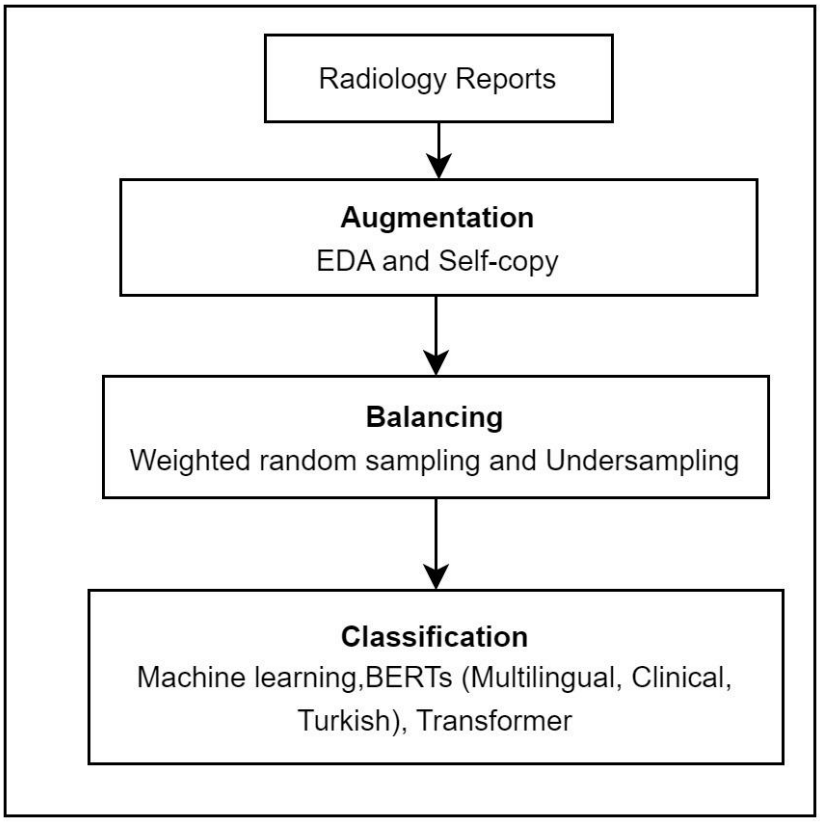


Figure 3.12. The workflow of the pre-processing methods for radiology reports.

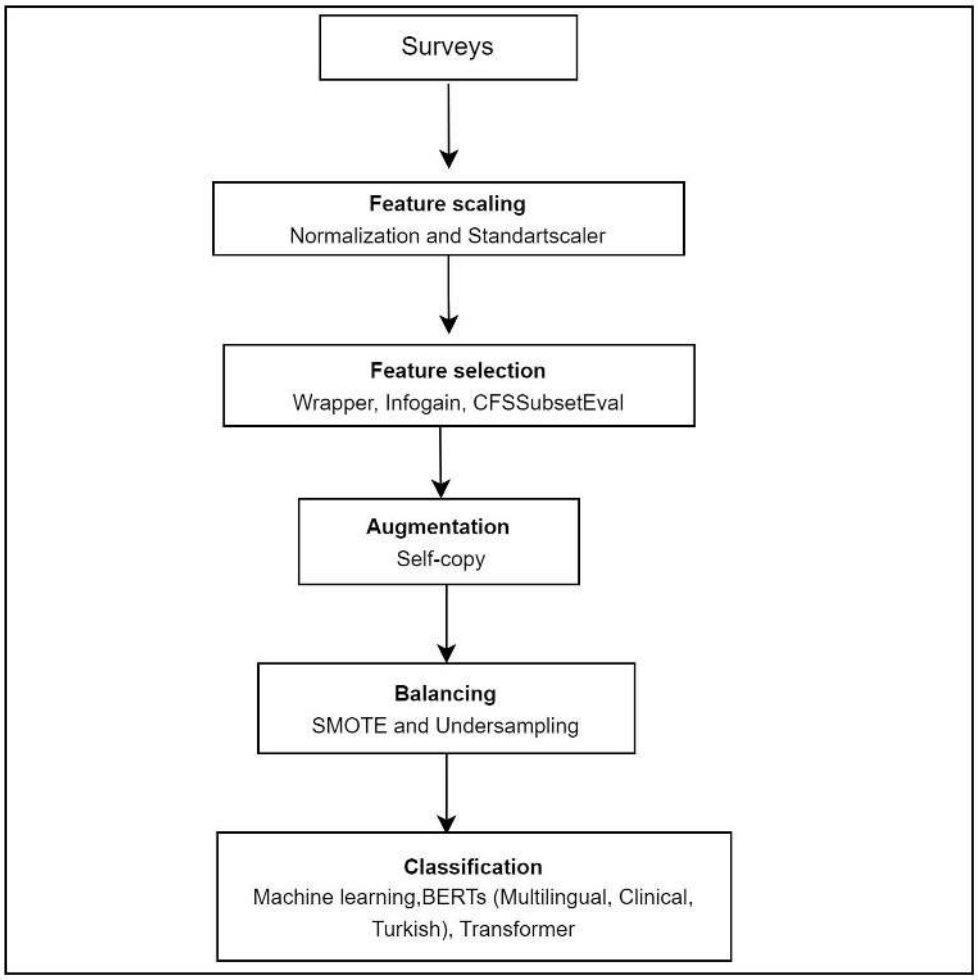


Figure 3.13. The worklow of the preprocessing methods for surveys.

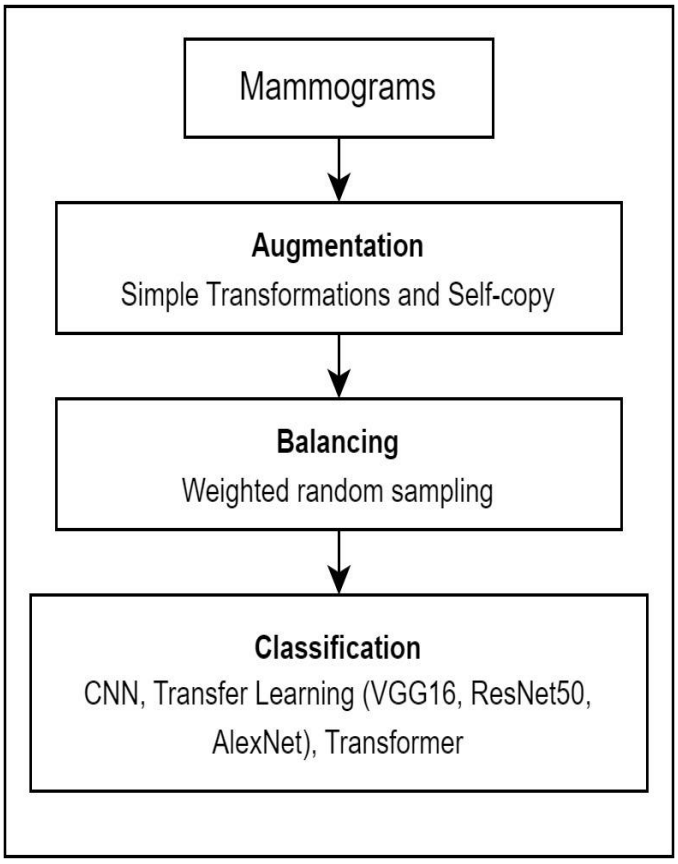


Figure 3.14. The workflow of the preprocessing methods for mammograms.

For all data types, first we formed the train-test datasets, after that we applied preprocessing.

As shown in the figures we applied preprocessing to train sets of each dataset. For the radiology reports dataset (in Fig 3.14) reports are first augmented by using EDA and self-copy methods, after that class balancing is applied.

As in Fig 3.15, for the survey dataset, first normalization methods are applied to the numeric features, after that feature selection, augmentation using self-copy, class-balancing using SMOTE and undersampling are applied.

For the mammogram image dataset (in Fig 3.16) augmentation using simple transformation and self-copy, after that class balancing using weighted random sampling are done.

Each pre-processing method according to applied data type is detailed below. We also show the table of the number of samples obtained as a result of the pre-processing method applied.

3.2.1.1. Augmentation

To achieve high accuracy in classification problems, the amount of data is important. Collecting data such as medical reports is usually tedious. Augmentation methodssuch as Easy Data Augmentation (EDA), self-copy and simple transformationsare used to find a solution to the small data problem.In this thesis, the self-copy and EDA methods for text data, the self-copy method for survey data, and simple transformations using torchvision transformations for image data are used to show the impact of augmentations on the success of thecalssification models. The number of samples in training in the "Survey," "Report," and "Image" datasets without the augmentation process are shown in Table 3.5.

Table 3.5. The number of training samples in the survey, report, and image datasets before augmentation.

	# of samples in the original training datasets		
Class labels	**survey**	**report**	**image**
Malignant	118	34	34
Benign	39	16	16
Total	157	50	50

Self-copy:

In this part, we increased the number of data instances by copying of each data sample in the training dataset. For the radiological reports dataset, each report in the benign class of the training dataset is copied by 26 times and each report in the malignant class of the training dataset is copied by 12 times. For the survey

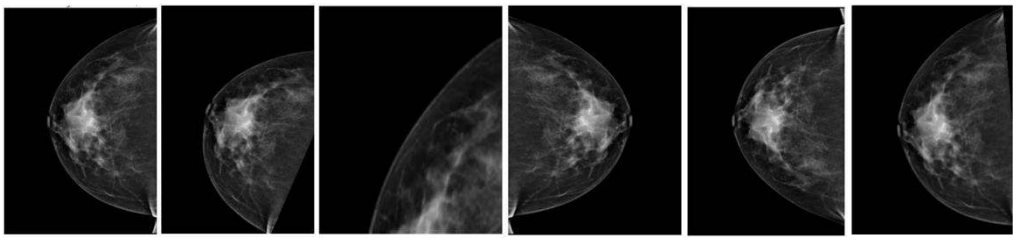
dataset, each survey in the malignant class and benign class of the training dataset is copied by 10 times. Table 3.6 shows the number of samples in the survey and report data after augmenting the data with self-copy method.

Table 3.6. The number of training samples in the report and survey dataset with self-copy.

# of samples in the augmented dataset with self-copy		
Classlabels	**report**	**Survey**
Malignant	408	1180
Benign	416	390
Total	824	1570

Simple transformations:

To enhance the image data, we use simple transformations, namely random rotation, random resized crop, random horizontal rotation, and random vertical rotation. In random rotation, the image is rotated by a random angle. In random resizedcrop, transformation crops an image at a random location and then resizes it to a specific size. In random horizontal rotation and random vertical rotation, an image is rotated horizontally and vertically respectively with a certain probability. Fig. 3.15shows the original image (on the left) and augmented images. As shown in Table 3.7, for each mammogram, 5 new samples are created and as a result, 96 benign and 204 malignant images are obtained.



a b c d e f

Figure 3.15. (a) Original, (b) random rotation, (c) random resized crop, (d) random horizontal rotation, (e) random vertical rotation, (f) random rotation.

Table 3.7. The number of traininginstancesin the augmented image dataset with simple transformations.

# of samples in the augmented dataset with simple transformations	
Class labels	**image**
Malignant	204
Benign	96
Total	300

EDA:

EDA (Wei and Zou, 2019) augments texts in several ways. These are random insertion, random deletion, random interchange and synonymous substitution. In random insertion, random words, apart from stop words, are selected and each word is replaced with a synonym of the word. This is repeated *n* times. For each text, *n* is the number of additions. In random deletion, each word is removed with random probability. In random swap, two words are randomly selected and their positions are swapped. In synonym replacement, *n* words except stop words are randomly selected and each word is replaced with a synonym of the word. In this part, Turkish WordNet (Çetinoğlu et al., 2018) is used as synonym dictionary when synonym replacement and random insertion are applied. The EDA technique of random insertion, random deletion, random swap, and substitution of synonyms is applied to each radiology report in the original dataset to increase the number of samples in the dataset. Using these methods, 12 new report samples were obtained from each benign and malignant report. We chose to heuristically generate 12 new reports for each sample according to the parameter recommendations of (Wei and Zou, 2019) for small training sets.Table 3.8 shows the number of samples in the training section after augmenting the data with EDA. Fig. 3.16 shows an original radiology report and augmented reports using random insertion, deletion, swap, and synonym replacement operations of EDA.

Table 3.8. The number of training samples in the augmented report dataset with EDA.

# of samples in the augmented dataset with EDA	
Class labels	**Report**
Malignant	408
Benign	192
Total	600

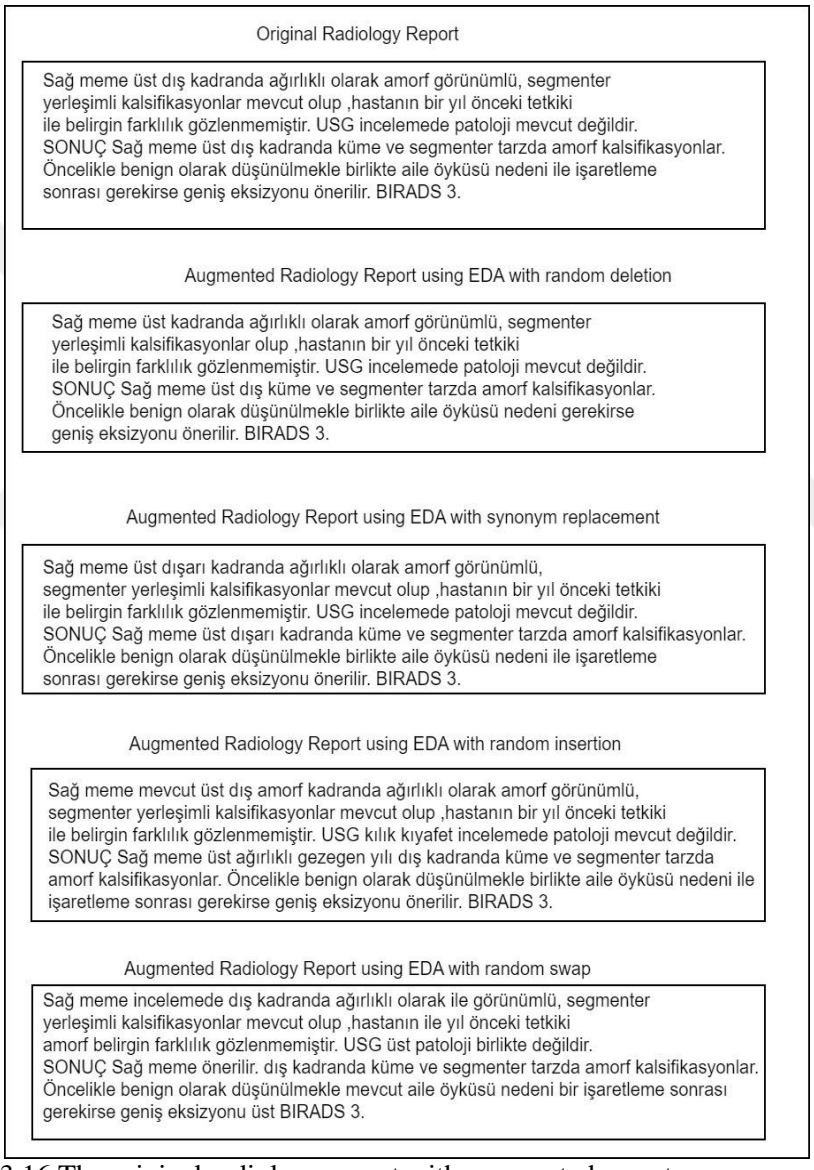


Figure 3.16.The original radiology report with augmented reports.

3.2.1.2. Class Balancing

Our collected data includes 20 benign and 42 malignant radiological reports and mammogram images as well as 126 malignant and 43 benign surveys. Therefore, the number of benign texts is much lower than the number of malignant texts, mammogramsand surveys. Such datasets are called imbalanced datasets, which may have a negative impact on minority class classification. We experimented with subsampling, weighted random sampling, and SMOTE techniques to balance the dataset and overcome this challenge.

SMOTE:

Synthetic Minority Oversampling Technique (SMOTE) is a well-known oversampling technique. This algorithm performs oversampling of the minority classes by creating synthetic instances to replicate the minority classes and increase their number of instances in the training set (Rao &Makkithaya, 2017; Rajendran et al., 2020). A random sample x is first selected from the minority class. Then k nearest neighbours are found for this sample, and y is randomly selected from these neighbours, and a synthetic sample is generated at a random point between the two examples x and y in the feature space. Since the new minority instances are generated by interpolating between multiple matching minority instances, the problem of overfitting is avoided (Rajendran et al., 2020; Batista et al., 2004).

After enriching the training dataset with survey, report, and image data, the number of samples in the malignant and benign classes is unbalanced. When we enrich the survey data with self-copies, we get unbalanced data, namely 1180 samples in the malignant class and 390 in the benign class. Thus, according to the algorithm SMOTE, as demonstrated in Table 3.9, we have a total of 1180 data instances, of which 1180 are malignant studies and 1180 are benign studies. Table 3.9 also presents the expanded data where only the algorithm SMOTE was used without increasing the number of malignant and benign studies. We have 118

malignant studies and 39 benign studies. When we match the survey data with SMOTE, we get 118 malignant studies and 118 benign studies.

Table 3.9. The number of training samples using SMOTE.

# of training samples of survey dataset				
	SMOTE		**without SMOTE**	
Class labels	**augmented with self-copy**	**without augmentation**	**augmented with self-copy**	**without augmentation**
Malignant	1180	118	1180	118
Benign	1180	118	390	39
Total	2360	236	1570	157

Weighted Random Sampling:

When we enrich the report data with the EDA method, we get 408 samples in the malignant class and 192 in the benign class. When we enrich the image data with transformations, we obtain 204 samples in the malignant class and 96 samples in the benign class. To balance these training samples of image and report data, we experimented with the weighted random sampling (WRS) technique (Efraimidis, 2015). First we counted the instancesin each class, then we calculated the associated weight for each class by taking the reciprocal of the number of samples belonging to each class, and finally we assigned the appropriate weight to each sample based on its class. Thus, the weight of the samples is used for the probability of selecting each sample.Table 3.10 shows the number of training samples in each class after balancing with weighted random sampling.

Table 3.10. The number of training samples using weighted random sampling.

	# of training samples			
	image with transformations		**report with EDA**	
Class labels	**WRS**	**Without WRS**	**Without WRS**	**WRS**
Malignant	204	204	408	408
Benign	204	96	192	408
Total	408	300	600	816

Table 3.11. The number of training samples using under sampling.

	#of training samples			
Class labels	**survey with self-copy**		**report with self-copy**	
	Under sampling	**Without under sampling**	**Under sampling**	**Without under sampling**
Malignant	390	1180	408	408
Benign	390	390	408	416
Total	780	1570	816	824

Undersampling:

If we use the self-copy method to augment the report dataset, we get 408 samples of the malignant class and 416 of the benign class. If we increase the number of training samples in the survey dataset using self-copy, we get 390 benign surveys and 1180 malignant surveys. To balance these training samples of report and survey data, we use the undersampling method. Undersampling randomly removes some malignant samples from survey samples, which are supplemented with self-copies, and some benign samples from report samples, which are supplemented with self-copies. The result is that after subsampling, we have samples in 408 benign and 408 malignant classes for report data and in 390 malignant and 390 benign classes for survey data, as shown in Table 3.11.

3.2.1.3. Preprocessing methods for survey data

The below subsections summarize the applied preprocessing techniques for only survey data.

Feature Scaling:

Most machine learning algorithms use the Euclidean formula in their calculations, which measures the distance between two data points, one of the distance metrics. To ensure that all values contribute equally, the values must be brought to the same unit. Data that varies in size increases the variance and does not have an equal effect in the distance calculations because it outweighs the low-value features in the weight calculations. When we examine our survey, it consists

of various data such as age, height, weight, number of births, and alcohol consumption. We have also seen that the height data takes values in the range of 140-180, while the number of births data takes values between 0-2. In this case, the model is built with the idea that the height variable is more important than the birth variable. Therefore, we decide that feature scaling should be performed on our data since there are different distributions for each feature in our data. The survey data is scaled by feature normalization and standardization as explained below.

- Normalization:

It scales each value in the dataset according to the min-max values in the corresponding column. We have brought all the attribute values in our survey to 0-1 range using min-max normalization method. This is done by importing and using the MinMaxScaler function from the Sklearn.preprocessing library in Python. Equation 3.5 shows the formula to calculate min-max normalized values.

$$x_{scaled} = \frac{x - x_{min}}{x_{max}} * (max - min) + min \quad (3.5)$$

Where x is the value to be normalized; $x_{min}$and x_{max} are the min and max values of the x value, max and min are the max and min values of the new range.

- Standard scaler:

It translates the variables into a distribution with a mean of 0 and a standard deviation of 1. It is found by subtracting the corresponding column mean from all the data in the dataset and dividing by the column standard deviation. Thus, all the observation units in the dataset get values between - 1 and 1. For this process, the StandardScaler method from Sklearn.preprocessing library is imported

into Python and used. Equation 3.6 shows the formula to find the standardized values.

$$y_{standardized} = \frac{(x-\mu)}{std.deviation} \quad (3.6)$$

Where the x is the value to be standardized and μis mean which is calculated by Equation 3.7:

$$\mu = {sum(x)}\Big/{count(x)} \quad (3.7)$$

And the $std.deviation$ is the standart deviation of the attribute and it is calculated by Equation 3.8:

$$std.deviation = \sqrt{sum((x-mean)^2)/count(x))} \quad (3.8)$$

Feature Selector:

Our survey dataset consists of 56 features. We select some of the features from the augmented and normalized data using feature selection algorithms available in the data mining tool WEKA. We select attributes from the training and test data using the Correlation-Based Feature Selection (CfsSubsetEval), Wrapper-Based Feature Selection (WrapperSubsetEval), and Information Gain-Based Feature Selection (InfoGainAttributeEval) methods.

(a) Correlation-based Feature Selection (CfsSubsetEval): The subset of features is ranked based on the correlation with the class label and other features by CfsSubsetEval. Subsets that have lower correlation with other features and high correlation with the class label are given a higher value. Moreover, this approach

does not consider relevant and redundant features of the dataset (Hall, 1999; Kumar et al., 2017).

(b) Wrapper: This method finds a subset of features that is searched in the feature space. The training process of a classifier is performed on this feature subset and the accuracy of the classifier is calculated to evaluate the success of this feature subset (Kumar et al., 2017).

(c) Information Gain Attribute Evaluation: In this method, information gain values of the attributes are computed according to the classification objective. Attributes are eliminated by setting a threshold value. If the information gain of the attributes is above the threshold, they are considered in the further processing steps (Kumar et al., 2017; Hall and Smith, 1998).

Principal Component Analysis (PCA):

Principal Component Analysis is used to reduce the dimensions of the data. PCA represents hidden parameters in the data to classify them and reduces the number of features in a dataset without losing the variability of the data (Ramadevi et al., 2015). In our work, we also use PCA to test the classification results. We determine the number of components and choose the k^{th} largest component that explains 89% of the total variability.

3.2.1.4. Text data pre-processing for classical machine learning models

Preprocessing methods are very important for text classification algorithms to improve data quality when using machine learning models. In this thesis, we apply the following preprocessing steps to the text reports using the Natural Language Tool Kit (NLTK) (Wang and Hu, 2021); a Python package that contains important preprocessing functions.

- *Lowercase conversion*: All uppercase letters are converted to lowercase otherwise uppercase and lowercase versions of the same

word are considered as different. For example, the word "Kitle" and "kitle", which means "mass" in English, are considered different even though they are the same. Therefore, converting each word to lowercase is one of the most important preprocessing steps.

- *Tokenization*: It is the process of separating each word in the whole document. In this step, each text is broken down into a list of individual words using NLTK-library. For example, the text "ciltaltıyağdokusuve meme başlarıdoğaldır" can be tokenized into "cilt", "altı", "yağ", "dokusu", "ve", "meme", "başları", "doğaldır".
- *Stop word elimination*: Stop words are words that are used frequently in a language (e.g., "a," "this," "that," etc.). One of the basic preprocessing steps is to remove stop words, because there are many words in Turkish that we repeat frequently in our daily life and they do not add value to the document for the classification process. Therefore, we first obtain a list of stop words for Turkish (Öztürk and Aksoy, 2018) and then remove them to obtain the less frequently used words that are more important for classification.
- *Converting numbers to words*: The textual documents used in our work contain numeric values expressed by using both digits and letters. For example, in some reports we have number 8 written as 8 (by using only digits), in some other documents this number is written by using word "eight". Actually, both terms have the same meaning, but when we tokenize the document, we store "8" and "eight" as different tokens. To solve this problem, we convert all the digit numbers into their equivalent word forms.
- *Punctuation elimination*: Text documents contain various symbols to separate sentences in the document. These symbols are unnecessary for text classification. All punctuation marks are removed from documents to improve the overall performance of the classification

algorithms. For example, "Cilt, cilt altı yağ dokusu ve meme başları doğaldır. Lipomatö parankim mevcuttur." After removing punctuation marks, we obtain the following text "Cilt cilt altı yağ dokusu ve meme başları doğaldır Lipomatö parankim mevcuttur."

- *Stemming*: Stemming is the reduction of words to their word stems, such as "boyutunda" and "boyutta," which are all based on the root word "boyut." Zemberek (Ataman and Özgüner, 2021), a natural language processing tool for Turkish, is used for stemming in this research. The Turkish stemmer removes the suffixes/affixes from words.

Preprocessing applied to an example sentence is given in Table 3.12.

Table 3.12. Output of preprocessing steps for sentence "Sol saat 8 de 5x9 mm boyutunda benign kriterler gösteren lezyonun takip ve kontrolü önerilir."

State	**Sentence**
Original	Sol saat 8 de 5x9 mm boyutunda benign kriterler gösteren lezyonun takip ve kontrolü önerilir.
Lowercase	sol saat 8 de 5x9 mm boyutunda benign kriterler gösteren lezyonun takip ve kontrolü önerilir.
Tokenization	sol saat 8 de 5x9 mm boyutunda benign kriterler gösteren lezyonun takip ve kontrolü önerilir.
Stop word elimination	sol saat 8 5x9 mm boyutunda benign kriterler gösteren lezyonun takip kontrolü önerilir.
Converting numbers to words	sol saat sekiz beşxdokuz mm boyutunda benign kriterler gösteren lezyonun takip kontrolü önerilir.
Punctuation elimination	sol saat sekiz beşxdokuz mm boyutunda benign kriterler gösteren lezyonun takip kontrolü önerilir
Stemming	sol saat sekiz beşxdokuz mm boyut benign kriter göster lezyon takip kontrol öner

3.2.1.5. Framework

In the survey part and text part to test machine learning models, we use the Waikato Environment of Knowledge Analysis (WEKA) tool developed by the College of Waikato (Wittenetal., 2005). This tool is open source and was developed based on JAVA. Using WEKA we can easily perform data preprocessing, classification, clustering, and association, regression, and feature selection methods. In the text part and the image part of the thesis, we use google colab to test the pre-trained (VGG16, ResNet50, AlexNet, and BERTs) and transformer models. Google colab is an interactive, fully cloud-based, easy-to-use and collaborative programming environment.

For all experiments, training, validation, and testing were performed using PyTorch (Paszke et al., 2019) version 1.8.1 with CUDA 10.2, using NVIDIA Tesla P100 GPUs for the larger models.

3.2.2. Classical Machine Learning Methods

In this thesis, we use classical models such as SVM, Random Forest, Bagging, etc. to classify radiology reports and surveys and compare their performances. How to represent texts in a collection is one of the most critical questions in text classification tasks. There are several techniques for converting texts into numerical vectors. In this work, we use Term Frequency - Inverse Document Frequency (TF-IDF) to obtain features from our created radiology reports dataset and use them for applying classical machine learning algorithms to classify the breast cancer radiology reports as malignant or benign. TF-IDF is one of the most commonly used methods in Text Mining. This method calculates the weights of words to show how significant a word is to the document. The formula for TF-IDF is shown in Equation 3.9.

$$tfidf(i,j) = tf(i,j)\log(N/df_i) \quad (3.9)$$

Where *i* is a word, *j* is a document, *tf (i, j)* is the number of occurrences of word *i* in document *j*, df_i is the number of documents containing word *i*, and *N* is the total number of documents.

a) SVM:

Support Vector Machines (SVM) is a supervised machine learning algorithm designed for classification and regression. SVM divides the data into two or more classes to find the optimal linear or nonlinear frontier (Agarwal, 2013). The most commonly used SVM classifier is a binary one. It tries to predict the class of the test sample between two possible classifications. It works as follows: One takes a certain number of features with the class label and tries to find the optimal linear separation hyperplane in an N-dimensional space, where N defines the number of features. Support vectors can be used to determine the maximum distance between the two classes (Soui et al., 2021). The best hyperplane is determined by dividing the maximum distance into two parts, and the class of new samples is determined according to which side of the hyperplane the new sample falls.

b) Random Forest:

It is commonly used in supervised machine learning tasks. It groups a set of decision trees that form a forest (Ao et al., 2019). Bagging, making predictions by generating randomness by repeatedly creating a single tree with the bootstrap sampling method, is the basis of random forests. Random forests are created using the bootstrap aggregation method. The observations for the trees are selected using the bootstrap random sampling method, and the variables are selected using the random subspace method. The predictions of each decision tree based on randomly selected data samples are used, and the result of the prediction with the most votes is selected as the final prediction. The larger the number of trees in the forest, the

better the prediction results. Each decision tree is a set of rules based on the values obtained from the input features (Agarwal, 2013).

c) Naïve Bayes:

The Naïve Bayes algorithm is one of the most common supervised machine learning algorithms and is based on Bayes Theorem. The Naive Bayes classifier calculates the posterior probability of each class and the class with the highest posterior probability is the prediction result. It is very successful in medical data applications. The Naïve Bayes classifier has a number of advantages, e.g., it is efficient in training, unaffected by unrelated characteristics, and able to process true discrete data (Maysanjaya et al., 2018).

d) Bayes Net:

Bayesian network is a successful graphical method used to show probabilistic dependencies between random variables. It identifies probabilistic relationships between variables associated with an outcome of interest. A Bayesian network consists of an acyclic direct graph and a set of conditional probability distributions. For each node of an acyclic directed graph, a random variable is represented. The connections between nodes indicate probabilistic dependence. The probability of each event of interest is calculated for each class and then assigned to the class with the highest probability (Choi et al., 2009).

e) C4.5 (J48):

This is a supervised learning algorithm that is based on decision tree model. It divides each part of the training data into smaller subsets to make a decision based on them. The normalized information gain that outputs the division of the data by the choice of an attribute is studied. The predictor variable with the highest information gain is identified, and the tree branches from this variable. In this way, the data is evenly distributed among the branches. After the first predictor variable is determined, the same process is repeated, this time using the information value of this determined predictor variable instead of the total entropy and calculating which of the remaining predictor variables provides a greater

information gain by dividing this determined variable. This process continues until all predictor variables are inserted into the tree (Venkatesan and Velmurugan, 2015).

f) Decision Stump:

It is a one-stage decision tree with a root and its leaves. A decision stump makes a prediction for each feature value. A threshold value is chosen for each continuous feature. The tree stump contains two leaves, one for values above the threshold and one for values below the threshold. For nominal features, a tree stump can be created as a leaf for each feature value, and these nodes are called intermediate nodes. Thus, each intermediate node is assigned to the class that represents the majority of the data. The prediction is made for each intermediate node usingthe majority of the data class (Decision stump, 2021).

g) Attribute Selected Classifier:

It selects the attributes using only the training data, then trains the classifier and evaluates it using the test data. First, the dimension of the training and test data is reduced to perform the classification task (Attribute selected classifier).

h) DeepLearning4j:

It is a deep learning package in WEKA. It includes convolution layers, dense layer, subsampling layer, batch normalization, LSTM, global pooling layer, and output layer to build neural network architectures.

Ensemble Learning Methods:

a) Bagging:

Bagging, that is, Boostrap is one of the most commonly used ensemble learning algorithms. It combines the results of multiple models to obtain a generalized result.

b) Adaboost:

The main idea behind Adaboost is to combine multiple weak learners into one strong learner to improve the performance of the classification model (Singer

and Marudi, 2020). The weak learner is generated with equal weights for each instance and trained with weighted data. The weights are increased when the items are misclassified; the weights of the correctly classified items are decreased. Then the weak learning algorithms are run again to obtain a weak classifier for the new weighted data. In this way, the weak learner output is trained to be the input to the other learner, and finally the results are combined to form the final decision limits (Agarwal, 2013).

3.2.3. Transformers

In order to use transformers, there must be a relationship between the characteristics, so survey data are not suitable for a transformer model. Therefore, transformers are only used for mammograms and radiology reportsin this thesis.
Transformers are kind of neural network architecture and introduced by Vaswani et al. (2017). Transformer model gains interest due to the state-of-the-art performance in NLP and Computer Vision tasks. Transformer is composed of encoder-decoder architectures that process sequential data in parallel without any recurrent network. The encoder architecture part of the transformer extracts information from the entire sequence, rather than looking at the last encoder state, as is common with RNN. Thus, for a given input element, each element of the output is assigned a higher weight by the decoder. In this study, we propose Transformer models that utilize the encoder module of the Transformer to perform classification by mapping given data to the image or text label.

3.2.3.1 Transformer for Mammograms

The objective of the Image Transformer is to learn the mapping from the sequence of image patches to the corresponding mammogram label. We use the mammogram and corresponding output in the Image Transformer. The architecture of the Image Transformer model is shown in Fig. 3.18. The proposed model is

composed of an embedding layer, an encoder and a classification layer. As first step, we divide the given image X with size $c \times h \times w$ where (h is the height, w is the width and c is the number of channels) into non-overlapping patches x_1, $x_2 \cdots$, x_n. Each patch is used as individual token of Transformer model (The patch size p is chosen as 16 × 16, 32 × 32 or 56 × 56, where a smaller patch size results in a longer sequence).

The patches are mapped to their embeddings with a learned embedding matrix E and the positional information $Epos$ is encoded and appended to the patch representation to keep the spatial arrangement of the patches. The result of the Embedding layer is $z_0$is passed to the Transformer encoder that is composed of L identical layers. At the last layer of the encoder, we take the average of the outputs and pass it to classification layer.

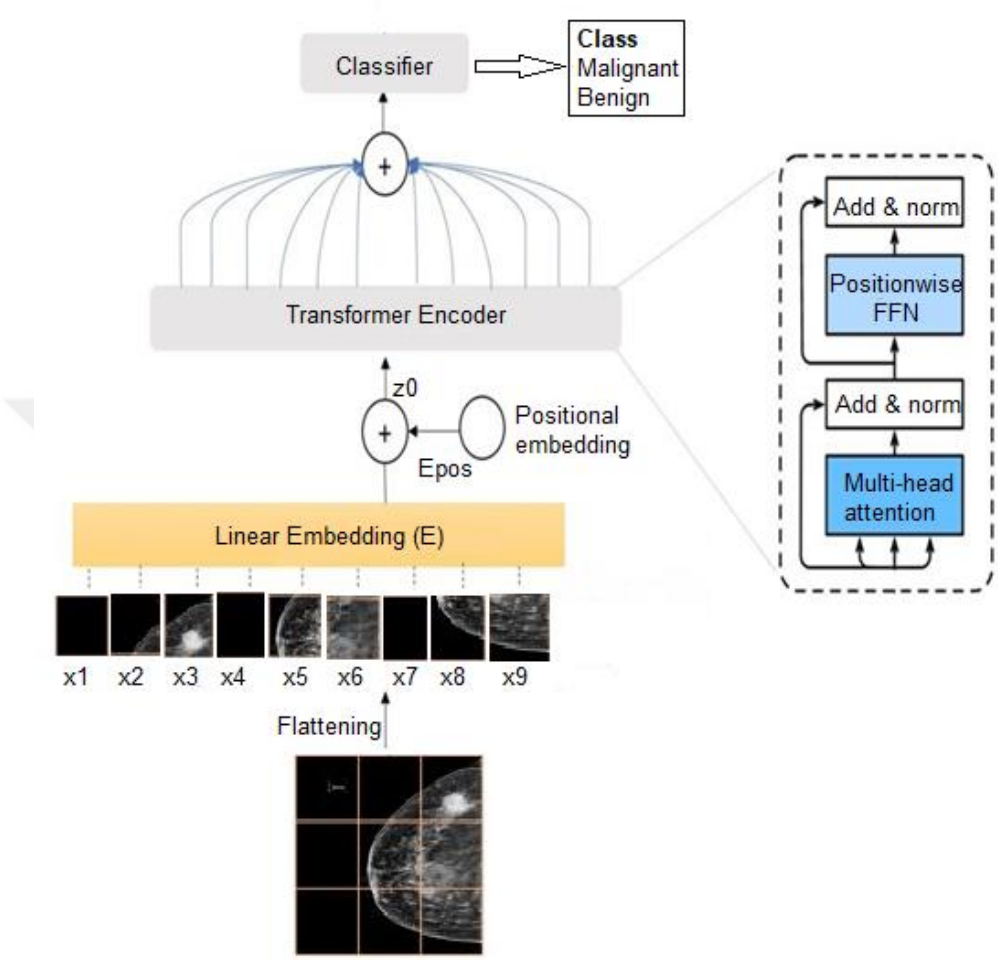


Figure 3.18. Overview of the Image Transformer architecture.

3.2.3.2 Transformer for Radiology Reports:

We use extracted text and corresponding radiology report label in Text Transformer. The architecture of the Text Transformer model is shown in Fig. 3.19.

The words $\{w_1, w_2, \cdots, w_n\}$ for text w_i are mapped to the corresponding embeddings in the embedding layer and the positional information *Epos* is encoded and appended to the text representation and fed into the encoder layer that is

composed of *L* identical layers. Classification layer takes the average of the output of the last encoder layer and returns label of the radiology report.

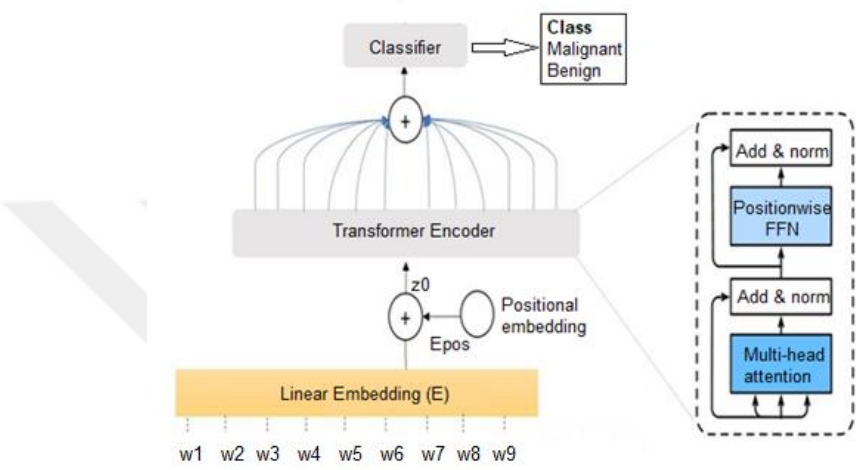


Figure 3.19. Overview of the Text Transformer architecture.

3.2.4. CNN for Mammograms

CNNs have been successfully used in BC detection. Fig. 3.20 shows a basic CNN for extracting information from mammograms. A CNN structure starts with the convolution blocks, denoted sequentially by conv1, conv2, and conv3, and ends with the global average pooling and the sigmoid layer. Fig. 3.20 shows the feature extraction using a mammogram as input matrix for convolution 1. All convolutional blocks contain maximal pooling and two convolutional layers.Two feature matrices are created with a 32 x 32 filter by convolution 1 and their output is used as input to convolution 2 to create feature matrices with a 64 x 64 filter. Just as before, the output of convolutional layer 2 is used as input to convolutional layer 3 to create feature matrices with a 32x32 filter. Then, the dimension of the feature matrices is reduced by applying maximum pooling.The output of convolution layer

3 is used as input for global average pooling to calculate the average of the feature matrices, and the calculated averages are passed to the next layer. And as the last step, a vector is created in the sigmoid layer. This vector consists of the probability of each label for each input and selects the label of the input that has the highest probability as the output.

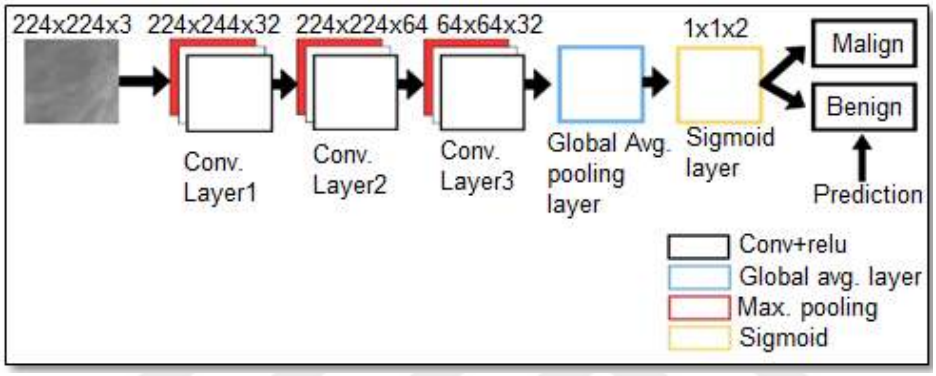


Figure 3.20.The architecture of CNN

3.2.5 Pre-trained models

Pre-trained models (VGG16, Resnet50, AlexNet, BERTs) are used to classify mammogram images and radiology reports. For survey data, we cannot use pre-trained models. The reason is that there must be a semantic relationship between features to use BERT, and VGG16, ResNet50, AlexNetare trained on images. Therefore, we cannot use pre-trained models for survey data.

3.2.5.1. Transfer Learning for Mammograms

Transfer learning is the reapplication of a previously trained model to a new problem. Pre-trained models were trained using the ImageNet dataset, which consists of 1000 categories. Most real-world problems usually have limited labeled data to train such complex models. To deal with this limited data problem, transfer learning is a very useful technique. In our work, we have investigated three

different network architectures for mammogram image classification. These transfer learning models are AlexNet, ResNet50, and VGG16. We chose these models because they are fast and provide higher accuracy in classifying mammograms. For all models, AlexNet, ResNet50, and VGG16, we used the same basic architectures as in the original work, as shown in Table3.13, Table3.14, and Table3.15, respectively.

Table 3.13.AlexNet Architecture.

Layer	**# of filters**	**Filter size**	**Stride**	**padding**	**Size of featıure map**
Input	-	-	-	-	227x227x3
Conv1	96	11x11	4	-	55x55x96
MaxPool1	-	3x3	2	-	27x27x96
Conv2	256	5x5	1	2	27x27x256
MaxPool2	-	3x3	2	-	13x13x256
Conv3	384	3x3	1	1	13x13x384
Conv4	384	3x3	1	1	13x13x384
Conv5	256	3x3	1	1	13x13x256
MaxPool3	-	3x3	2	-	6x6x256
Dropout1	0.5	-	-	-	6x6x256

Table 3.14. ResNet50 Architecture.

Layer-name	**Output size**	**50-layer**
Conv1	112x112	7x7,64, stride 2
Conv2.x	56x56	3x3 max pool, stride 2
	56x56	[1x1, 64]x3
		[3x3, 64]x3
		[1x1, 256]x3
Conv3.x	28x28	[1x1, 128]x4
		[3x3, 128]x4
		[1x1, 512]x4
Conv4.x	14x14	[1x1, 256]x6
		[3x3, 256]x6
		[1x1, 1024]x6
Conv5.x	7x7	[1x1, 512]x3
		[3x3, 512]x3
		[1x1, 2048]x3
	1x1	Average pool, 2-d dc, softmax

Table 3.15. VGG16 Architecture.

Layer	**type**	**Input**	**kernel**
Data	input	224x224x3	N/A
Conv1-1	convolution	224x224x64	3x3
Conv1-2	convolution	224x224x64	3x3
Pool1	max pooling	112x112x64	2x2
Conv2-1	convolution	112x112x128	3x3
Conv2-2	convolution	112x112x128	3x3
Pool2	Max pooling	56x56x128	2x2
Conv3-1	convolution	56x56x256	3x3
Conv3-2	convolution	56x56x256	3x3
Conv3-3	convolution	56x56x256	3x3
Pool3	Max pooling	28x28x256	2x2
Conv4-1	convolution	28x28x512	3x3
Conv4-2	convolution	28x28x512	3x3
Conv4-3	convolution	28x28x512	3x3
Pool4	Max pooling	14x14x512	2x2
Conv5-1	convolution	14x14x512	3x3
Conv5-2	convolution	14x14x512	3x3
Conv5-3	convolution	14x14x512	3x3
Pool5	Max pooling	7x7x512	2x2
Fc6	Fully connected	1x1x4096	1x1
Fc6	Fully connected	1x1x4096	1x1
Fc7	Fully connected	1x1x2	1x1

To classify a mammogram image as malignant or benign, were placed only the last layer of the AlexNet shown in Fig. 3.21 the VGG16 shown in Fig. 3.22, and the ResNet50 model shown in Fig. 3.23 with our new last layer. Thus, the new last layer of the pre-trained network has 2 categories instead of 1000 categories. We use AlexNet, VGG16,and ResNet50 as pre-trained networks by using their weights as initial weight values and training only the classifier with our dataset. As shown in Fig.3.21, Fig. 3.22, and Fig. 3.23, we freeze all layers except the last layer, add a new layer FC -2 with a soft-max activation function, and then re-train the last fully connected layer with fixed weights of the previous layers.The exception is the last fully-connected layer which number of nodes depend on the number of classes in the dataset. In our work, we implemented transfer learning architectures using pytorch and changed the last fully-connected layer to output 2 classes. We analyze the effect of initializing networks with pre-training on the

ImageNet dataset, and then fine-tuning on mammogram images. Fine-tuning is training a CNN using pre-trained weights. That fine-tuning method is very successful in improving the result of the model.

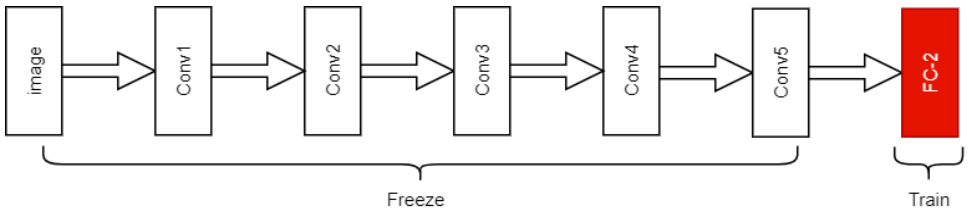


Figure 3.21.AlexNetto classify mammograms.

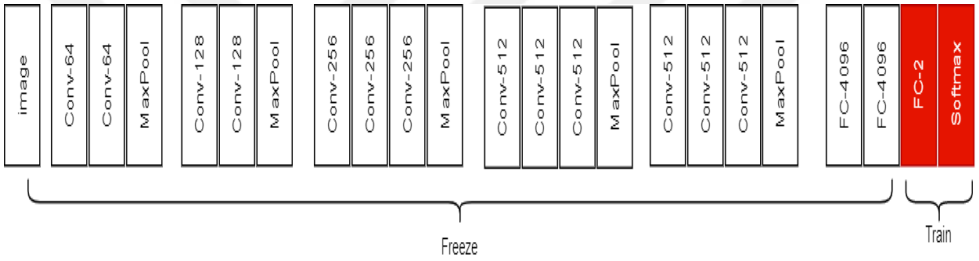


Figure 3.22.VGG16 to classify mammograms.

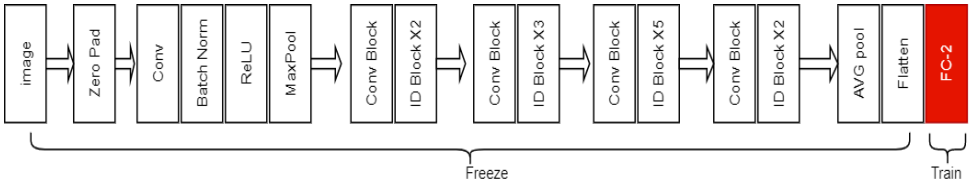


Figure 3.23.ResNet50 classify mammograms.

3.2.5.2. BERTs for Radiology Reports

In the NLP field, a large amount of training data is needed to work with Deep Learning-based models. Bidirectional Encoder Representations from Transformers (BERT) is a major turning point in the field of NLP to solve the problem of small training data in deep learning. Devlin et al. (2018) used two unsupervised techniques to train BERT, namely Masked Language Model (MLM)

and Next Sentence Prediction (NSP). In MLM, they randomly mask words in the sentence and then try to predict them. In NSP, given the first sentence the technique predicts if a chosen next sentence is probable or not. BERT uses both directions and the full context of the sentences to predict masked words. The overall architecture of the BERT (Devlin et al., 2018) model is given in Fig. 3.24. The basic BERT model contains an encoder with 12 Transformer blocks, 12 self-attention heads, and a hidden size of 768. The network takes as input a sequence of no more than 512 tokens and outputs the representation of the sequence. The sequence consists of one or two segments, where the first token of the sequence is always [CLS], which contains the special classification embedding, and another special token [SEP] is used to separate the segments (Tokgoz et al., 2021). The bidirectionality of the BERT model distinguishes it from previous language models. Google has provided pre-trained BERT models for various data sets. These models are available in various languages such as Turkish, Chinese, and multilingual versions. In this study, we evaluated and compared three different models based on BERT, namely BERT$clinical$, BERTT$urkish$ and BERT$multilingual$, to show the performance of the models with our dataset. For the radiology report classification task, we use the final hidden state of the first token [CLS] as the representation of the whole sequence.

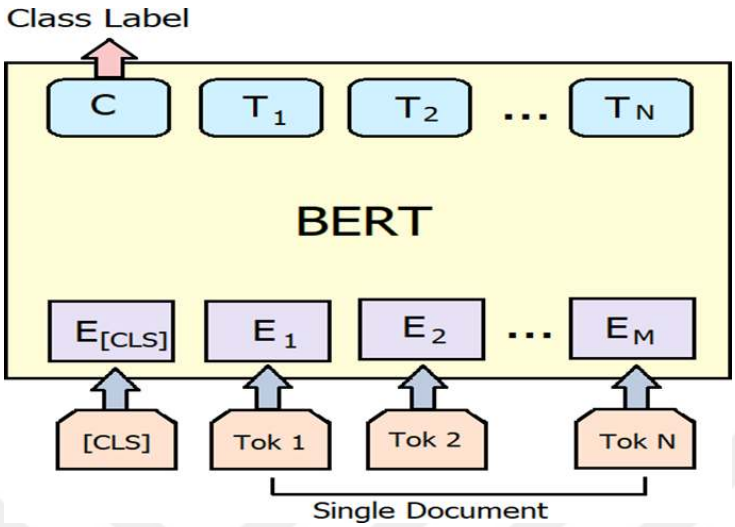


Figure 3.24. The architecture of BERT (Devlin et al., 2018) for created Turkish radiology report data.

Multilingual BERT model:

Multilingual BERT model is introduced by Devlin et al. in 2018. The multilingual BERT model is pre-trained on 104 languages, including Turkish. The training data consists of Wikipedia articles in each language. It does not use markers to indicate the input language and has no explicit mechanism to ensure that translation-equivalent pairs have similar representations. The multilingual model of BERT has two versions that differ in capitalization and punctuation. In this study, the case-sensitive version was used when the text data was not converted to lowercase.

Turkish BERT model:

We also used the Turkish BERT (Turkish Bert Model, 2020) model to classify Turkish radiology reports. The Turkish BERT model is the result of training with Turkish Wikipedia corpus and released by the MDZ Digital Library team. The current version of the model is trained on a filtered and sentence

segmented version of the Turkish OSCAR corpus, a recent Wikipedia dump, various OPUS corpora and a special corpus.

Clinical BERT model:

ClinicalBERT is a modified BERT model: Specifically, the representations are learned using medical notes and further processed for downstream clinical tasks. Clinical BERT is published by Alsentzer et al. (2019). The quality of learned representations of text depends on the text the model was trained on. BERT is pretrained on BooksCorpus and Wikipedia. The Clinical BERT model, which was specifically pre-trained for clinical domains using MIMIC notes, was used to see if experimental results improved when a specific domain model was used.

Distil-BERT model:

BERT (Devlin et al., 2018) is a very large and memory-hungry model that is slow in training and testing phases. Therefore, DistilBERT (Sanh et al., 2019) is proposed that is a “distilled" version of BERT, which is smaller and faster than BERT without reducing the accuracy of it. The general architecture is the same as that of BERT, except that token type embedding and the pooler have been removed, while the number of layers has been reduced by a factor of 2, which has a significant impact on computational efficiency. The model was distilled on very large stacks with dynamic masking and with prediction of the next sentence (NSP) (Verma, 2021). Some percentage of the input tokens are masked (Replaced with [MASK] token) at random and the model tries to predict these masked tokens. Masking and NSP here refer to the process of converting a word to be predicted in the Masked Language model to ["MASK"] and training the entire sequence to predict that particular word. The architecture of DistilBERT is given in Figure 3.25. As can be seen in Fig. 3.25, the architecture of DistilBERT consists of student and teacher networks and each network contains encoders (the block containing Attention, Normalization, Feed Forward Network (FFN), Normalization

placed on top of each other). Attention matrices generated by Multi-Head Attention (MHA) reduce the loss between Multi-Head Attention (teacher) and Multi-Head Attention (student). Similarly, Hidden States reduce the loss between Hidden State (Student) and Hidden State (Teacher) outputs of encoder stacks in Student Teacher.

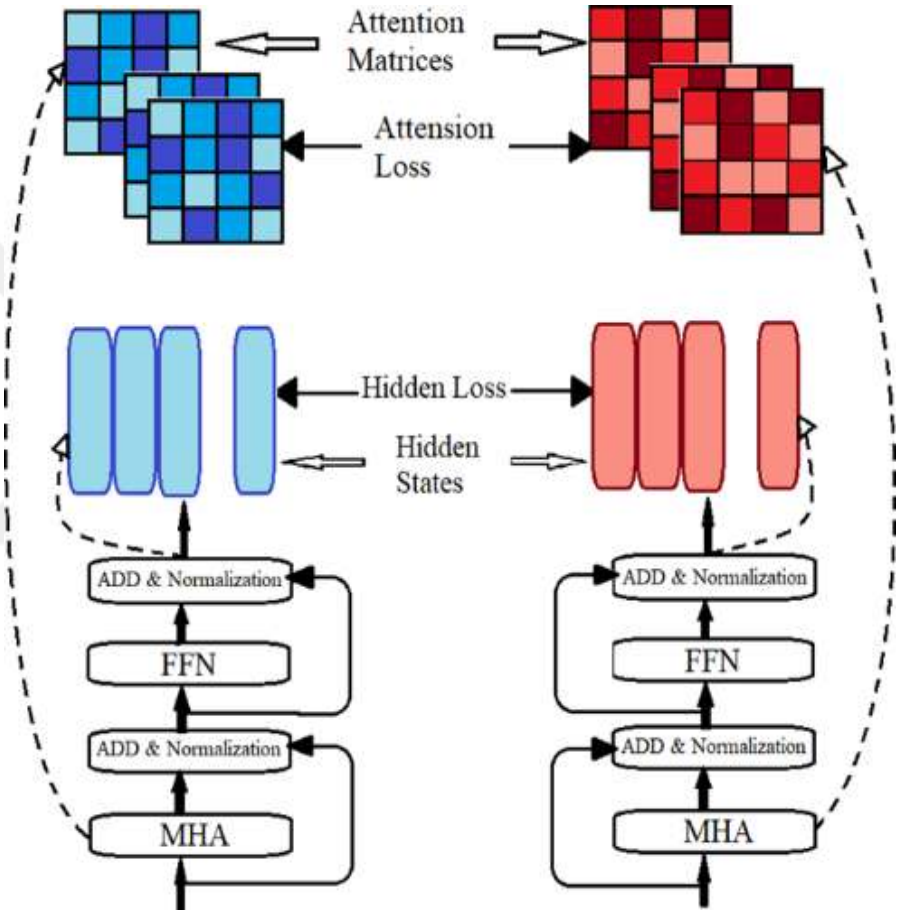


Figure 3.25. The architecture of DistilBERT (Sanh et al., 2019) for created Turkish radiology report data.

3.2.6. Proposed Methods

3.2.6.1. A Hybrid ofAll Three Data Types

In our thesis, we proposed to develop a novel hbyrid method for breast cancer prediction to reduce the cost of breast biopsies by considering traditional models and deep learning models such as Transformers, BERTs, VGG16, etc. using newly collected hospital datasets. The proposed model combines all three types of data for each patient to reduce the cost of breast biopsies. In other words, the proposed model takes newly collected survey, newly collected radiology report, and newly collected mammography data for each patient as input and generates an output. Our proposed model consists of a variety of models for each data type.

With the increasing success of deep learning methods in computer vision tasks, CNNs have been applied to many image classification problems, especially in medicine (Yadav and Jadhav, 2019). As a result, they have become very popular in breast cancer diagnosis (Alanazi et al., 2021). Deep learning methods require a large dataset to learn better during training. However, medical image datasets are of limited size. Therefore, applying the CNN method to solve medical image problems is a challenging task. Pretrained models are the most common technique to solve the problem of limited training data and have been used by researchers to analyze medical images. The most commonly used pre-trained models are ResNet (Chougrad et al., 2017), AlexNet (Hassan et al., 2020) and VGG16 (Falconi et al., 2020) on INbreast, CBIS-DDSM, etc. In this work, AlexNet, ResNet50 (He et al., 2016) and VGG16 (Simonyan and Zisserman, 2014) pre-trained models and Transformer models (Vaswani et al., 2017) are used to classify our newly created mammography dataset.

Another important data type is radiological reports, which are interpretations of diagnostic imaging for breast cancer prediction. It is very difficult to automatically extract meaningful data from radiology reports because they contain limited vocabulary and unstructured data (Suárez-Paniagua et al., 2021).

However, extracting information from radiology reports is very important for early detection of breast cancer. BERT and Transformer models are used to test our newly created Turkish radiology report dataset.

A survey to predict breast cancer is also a branch of this proposed model. A survey of patients with breast symptoms is conducted. Various machine learning algorithms (SVM, C4.5, AdaBoost, etc.) are used to classify our newly created dataset of Turkish surveys.

The most successful models with survey, mammography, and radiology report data are selected. Experimental results show that the pre-trained VGG16 model performs best for mammograms, the Transformer model with BERTmultilingual performs best for radiology reports, and Atttribute selected classifier with pure data performs best for surveys. The predictions of these models are combined by majority voting. The most predicted label is accepted as the prediction result of the hybrid model. Our hybrid model then provides a result on whether the patient is malignant or benign to reduce the cost of breast biopsy. An overview of the proposed model is shown in Fig. 3.26.

Similar to the literature, we used transfer learning and BERTs to solve the problem of limited data in our newly collected dataset. The main goal of this hybrid method is to combine all three types of data for each patient to reduce the cost of breast biopsies, depending on how much data we have about the patient. However, the literature typically uses only one type of data set to classify breast cancer, such as only image data, only radiology reports, or only survey data. Another novelty is that we collected a new dataset of Turkish radiology reports, Turkish surveys, and mammograms to test the performance of the proposed model. On the other hand, survey data have not been previously used in the literature to classify breast cancer. However, the proposed model is tested on newly created surveys with the help of an expert.

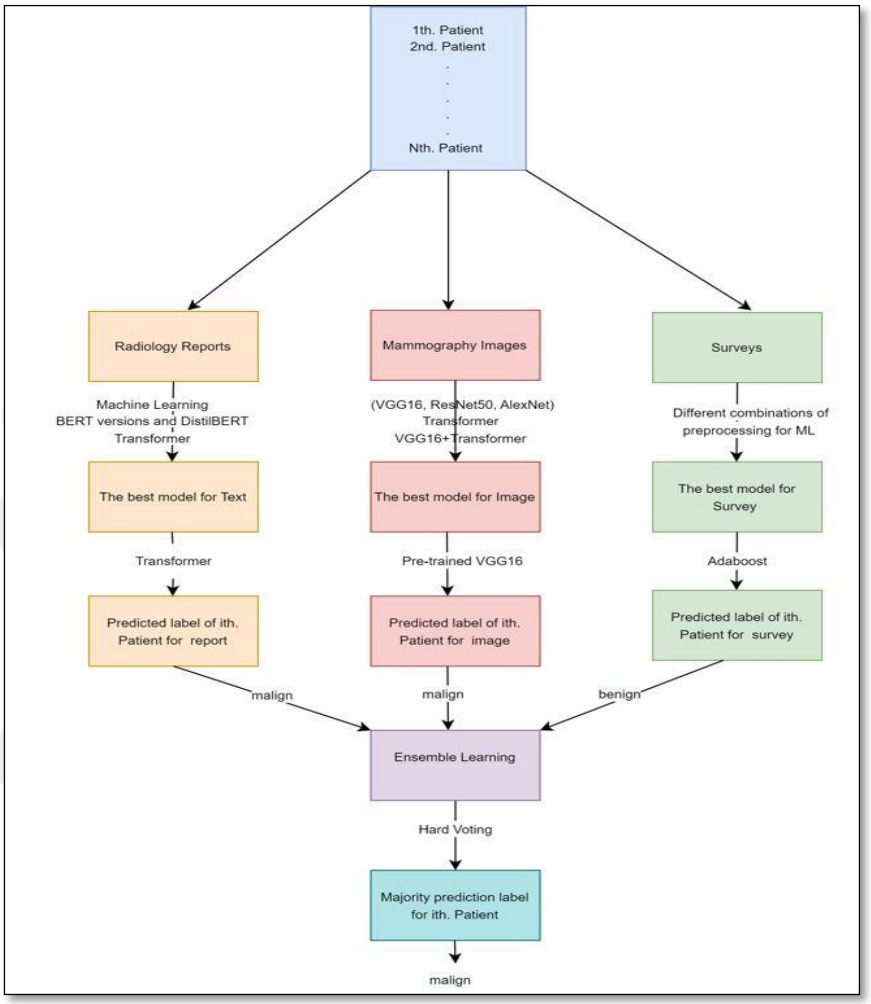


Figure 3.26.The workflow of the proposedsurvey+mammogram+report model.

3.2.6.2. A Risk-based Hybrid Neuro+Fuzzy Rule-based System

In this thesis, we also proposed a risk-based hybrid model consisting of neuro and fuzzy rules to calculate BC risk from mammograms. This proposed model uses Deep Learning (ResNet50 and CNN-based) and a rule-based fuzzy system to calculate the BC risk. This hybrid system creates a simple rule-based classifier to assign class labels based on the calculated risk scores. This hybrid model can be considered as a fuzzy system with a simple classifier at the end. In

this hybrid model, the deep learning part works as a preprocessor. It generates probabilities for each class from the training data for fuzzy rules or membership functions and then fades into the background. Then the fuzzy system and the classifier come into play, working as postprocessors in the hybrid model. The risk-based hybrid system is notable for health outside the classical BC detection models. Because the proposed model prevents unnecessary breast biopsies and their negative consequences such as costs, anxiety, etc., by capturing uncertain data well and expressing them in real numbers in breast cancer diagnosis. However, in the literature, there is no such study to classify mammogram images.

The experimental results show that using DL alone to classify mammograms leads to worse classification results than using the rule-based fuzzy system with DL models. Therefore, using our hybrid rule-based neuro+fuzzy model reduces the number of breast biopsies and diagnoses BC more accurately.

As shown in Fig. 3.27, our risk-based hybrid model architecture includes two distinct parts. The first part of the proposed model is a DNN model and the second part of the proposed model is a Fuzzy Rule Based System (FRBS). In the first part of the proposed system, a DNN model is trained using mammograms.

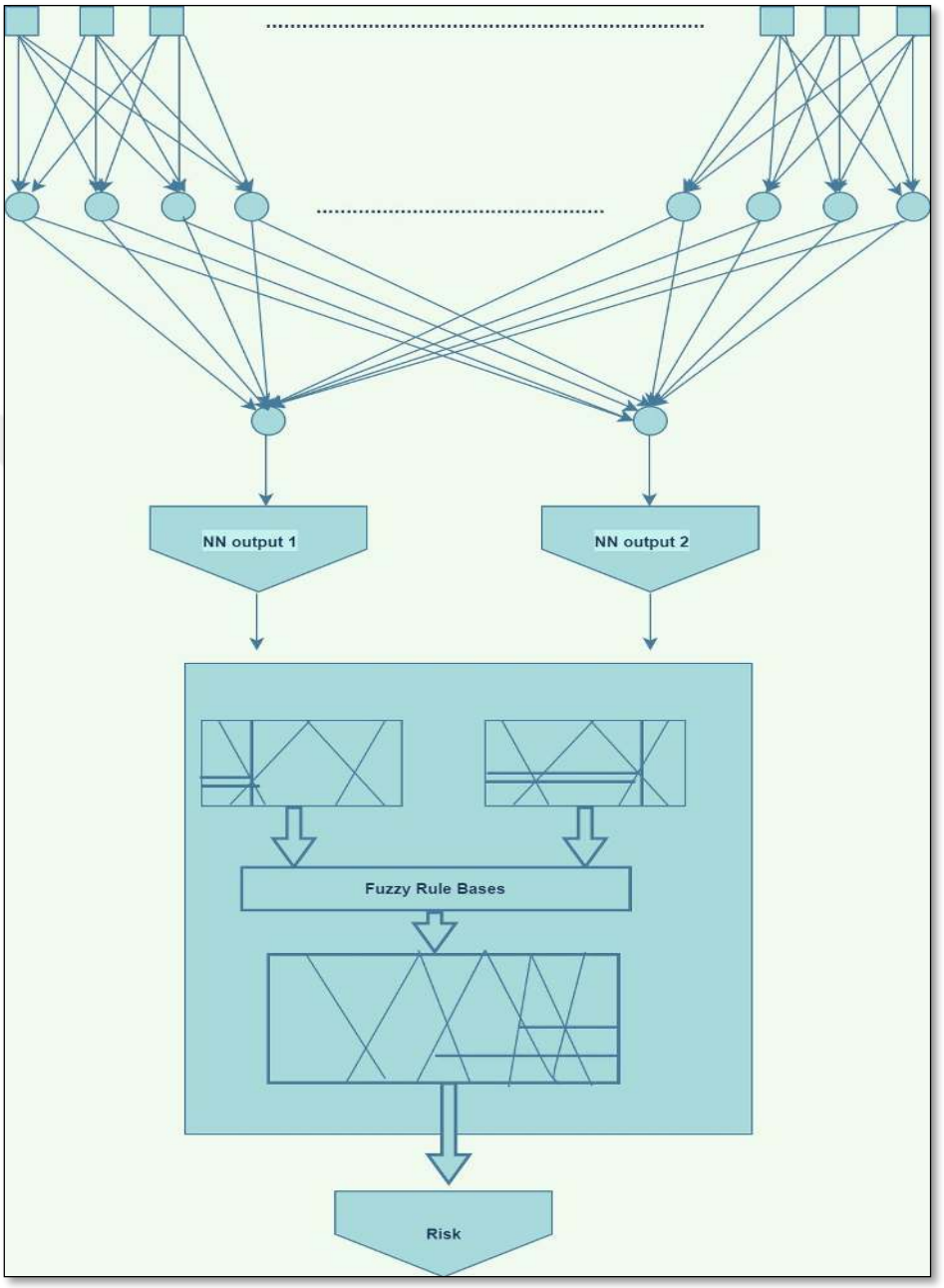


Figure 3.27. The workflow of the proposed neuro+fuzzy model.

From the DNN model, we obtain two output variables, namely NN-Output 1 and NN-Output 2. NN-Output 1 indicates the probability of mammogram images that are of the malignant type, while NN-Output 2 indicates the probability of mammogram images that are of the benign type. In the second phase, FRBS uses the output of DL as input to calculate the BC risk score using fuzzy methods.And then the mammography images are classified more accurately based on these risk scores.As an example, we can say that a person's risk for BC is high if the person's calculated score is 90% and therefore the person is classified as malignant on mammography.

In the proposed system, the CNN, which is used as a base, and the ResNet50 neural network are trained on the mammograms. To evaluate the performance of each network, only calcification type mammograms are considered.

Fuzzy approach:

Fuzzy logic (Zadeh, 1988) is based on fuzzy sets and subsets. In a fuzzy set (FS), each object has a degree of membership. The membership degree of the object can be any value in the range (0, 1) and is represented by the membership function *M(x).*

In this model, the FS is used to determine the risk values of mammograms that provide input to the rule-based classifier in the hybrid model. In the second phase of the hybrid model, the fuzzy methods are used, working as a post-processor. Experts can make a decision about an individual's BC risk score using our hybrid model. Two output variables, namely NN-Output 1 and NN-Output 2, generated by the first part of the hybrid model, are used as inputs by FRBS to generate the risk scores for the BC as output.

Table 3.16. The range of the probability that belongs to a malignant/benign type using the outputs of the DL part.

NN output (class label)	Probability of malign/benign class		
	Max	Min	MEAN
NN output 1 (Benign)	0.99	0.39	0.69
NN output 2 (Benign)	0.59	0.0001	0.30
NN output 2 (Malign)	0.99	0.53	0.76
NN output 1 (Malign)	0.38	0.008	0.19

First of all, we made a decision about the factors affecting BC risk scores and then described the output and input variables. The probabilities of the maximum and minimum for the benign and malignant types calculated in the first phase of the hybrid model are listed in Table 3.16 and are used to define the input and output variables in the FRBS. Looking at the third row in Table 3.16, we see that the result of NN-Output 2 for malignant-type mammography varies from "0.53" to "0.99". FRBS contains 2 inputs, namely the probability that the mammogram belongs to the benign type (NN-Output2) and the probability that it belongs to the malignant type (NN-Output1), and 1 output, namely the risk score.

FS takes values between "0 "and "1" as input and produces risk values between "0" and "100" as output. We chose trapezoidal and triangular MFs, which are commonly used and have simple shapes. Fig. 3.28 shows the use of both triangular and trapezoidal MFs to describe NN-Output 1 and NN-Output 2. Similar to Fig. 3.28, Fig. 3.29 shows the use of triangular and trapezoidal MFs to compute risk values. The input and output variables of FRBS are fuzzified using these MFs (Singh and Lone, 2020; Jang et al., 1997).

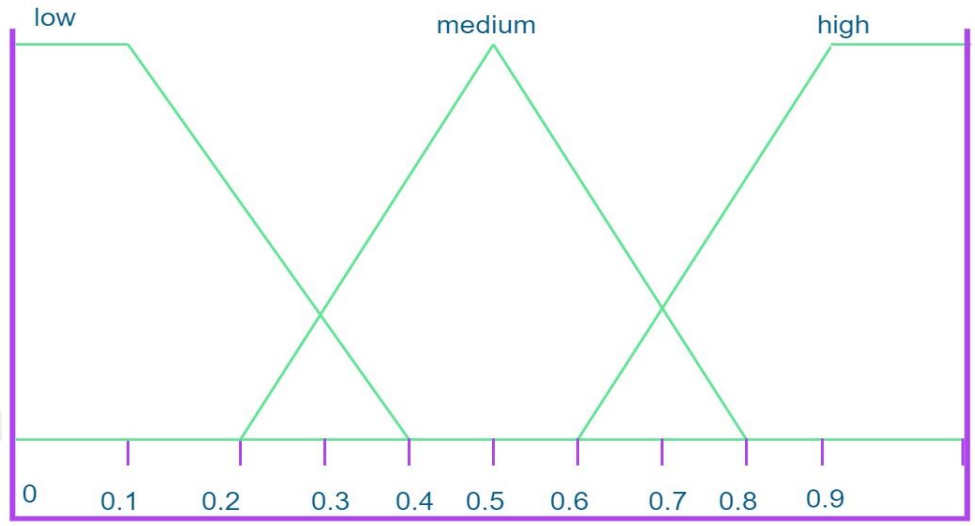


Figure 3.28. Membership functions of the input variables NN output 1 and NN output 2.

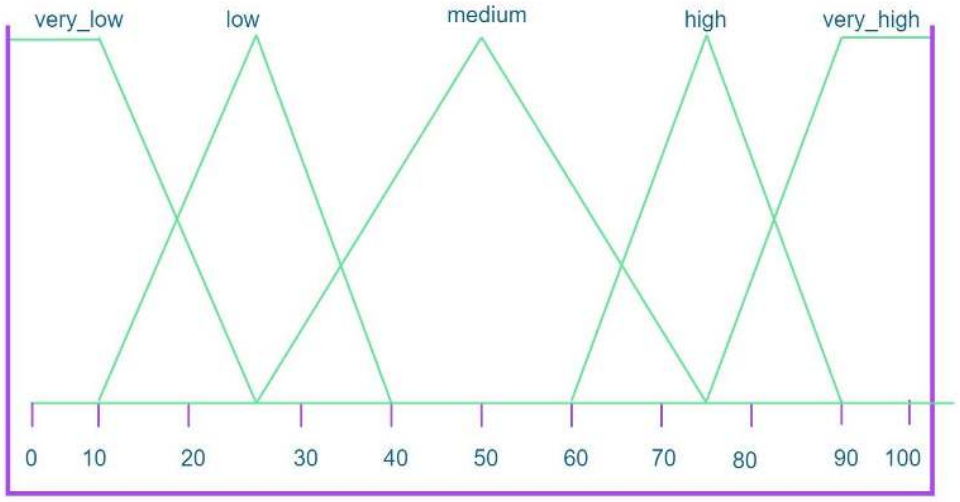


Figure 3.29. Membership functions of the output variable which is the BC risk.

The main idea behind FRBS is to avoid the negative consequences such as anxiety, cost, etc. of unnecessary biopsies with BC risk scores. First, we created proper ranges for each input and output variable, and then found and assigned the correct linguistic terms. A domain expert validated all of these ranges. Table 3.17 shows "*low*," "*medium*," and "*high*" linguistic terms with minimum and maximum values for benign and malignant class types.

Table 3.17. Distribution of probability values obtained from the neural network model.

Probability of each class	Min	Max	MEAN
NN output 1 (Benign) Low	0.1	0.4	0.25
NN output 1 (Benign) Medium	0.2	0.8	0.5
NN output 1 (Benign) High	0.6	0.9	0.75
NN output 2 (Malign) Low	0.1	0.4	0.25
NN output 2 (Malign) Medium	0.2	0.8	0.5
NN output 2 (Malign) High	0.6	0.9	0.75

As a result of training the DL part, we have two outputs, NN-Output 1 and NN-Output 2. The lower, upper, and middle bounds of NN-Output 2 are “0.1”, “0.9”, and “0.5”, respectively, as shown in Table 3.17. Thus, the *low* linguistic variable is defined between “0.1” and “0.4”, the *high* linguistic variable is defined between “0.6” and “0.9”, and the *medium* linguistic variable is defined between “0.2” and “0.8”. Some probabilities determined by neural networks may be smaller than “0.1” or larger than “0.9”. The probabilities those are smaller than “0.1” and the probabilities that are larger than “0.9” are included in the *low* and *high* linguistic variables, respectively. When the probability value is between 0 and 0.1 or between” 0.9” and “1”, the degree for *low* and *high* linguistic variables is “1”. When the probability value falls below “0.9”, the degree of high linguistic variable decreases. The probabilities between “0.2” and “0.8” are defined as medium linguistic variable using the triangular MF. The degree of medium linguistic variable is “1” for the probability value of “0.5”.

As can be seen in Fig. 3.27, the lower and upper bounds of the trapezoidal MF for the *low* linguistic variable are "0.1" and "0.4", respectively, and Equation 3.9. is used to compute the degree of the *low* linguistic varable.

$$\mu_{low} = \begin{cases} 0, & x > 0.4 \\ \frac{(0.4-x)}{0.3}, & 0.1 \leq x \leq 0.4 \\ 1, & x < 0.1 \end{cases} \quad (3.9)$$

In the same way, the lower and upper bounds of the trapezoidal MF for the *high* linguistic variable are "0.6" and "0.9", respectively, and Equation 3.10 is used to compute the degree of the *high* linguistic varable.

$$\mu_{high} = \begin{cases} 0, & x < 0.6 \\ \frac{(x-0.6)}{0.3}, & 0.6 \leq x \leq 0.9 \\ 1, & x > 0.9 \end{cases} \quad (3.10)$$

The lower and upper bounds of the triangular MF for the *medium* linguistic variable are "0.2" and "0.8", respectively and a value of *m*=0.5. The degree of the *medium* is computed using Equation 3.11.

$$\mu_{medium} = \begin{cases} 0, & x \leq 0.2,\ x \geq 0.8 \\ \frac{(x-0.2)}{m-0.2}, & 0.2 < x \leq m \\ \frac{0.8-x}{0.8-m}, & m < x < 0.8 \end{cases} \quad (3.11)$$

In addition, risk is defined in terms of the linguistic variables *very low*, *low*, *medium*, *high*, and *veryhigh*, as you can see in Fig. 3.28. The BC risk for *high* or *low* variables is defined as the average value for 50%. The values between "0" and "50 "(when the risk value is less than 50%) are explained by *very low* and *low*, while the values between "50" and "100" (when the risk value is more than 50%)

are defined by the linguistic variables *very high* and *high*, and the risk is expressed by the linguistic variable *medium* when the risk value is 50%. The ranges of risk values for the linguistic variables *very low, low, medium, high,* and *very high* can be found in Table 3.18. We also use triangular MFs for the risk values in the range of [10, 90] and trapezoidal MFs for the risk values in the range of [0, 10] and [90, 100].

Table 3.18 The ranges of risk values for linguistic variables.

Linguistic variables	**Mininum RS**	**Maximum RS**	**Average of RS**
Very low	0	10	5
Low	10	40	25
Medium	25	75	50
High	60	90	75
Very high	90	100	95

Calculation of Risk
If **benign** is low and **malign** is low, then **risk** is medium
If **benign** is low and **malign** is medium, then **risk** is high
If **benign** is low and **malign** is high, then **risk** is very high
If **benign** is medium and **malign** is low, then **risk** is low
If **benign** is medium and **malign** is medium, then **risk** is medium
If **benign** is medium and **malign** is high, then **risk** is high
If **benign** is high and **malign** is low, then **risk** is very low
If **benign** is high and **malign** is medium, then **risk** is low
If **benign** is high and **malign** is high, then **risk** is medium

Figure 3.30. Calculation of risk using fuzzy rules.

In the end, the algorithm called risk calculation is used to determine the linguistic variable of BC risk. Fig. 3.31 shows the algorithm which consists of nine fuzzy rules to calculate the BC risk value, and all of these rules are created by a doctor. To calculate the BC risk values using the rules of the algorithm, NN-Output 1and NN-Output 2 are given as input to the Fuzzy Inference System (FIS), and then the fuzzy risk is converted into numerical risk using the defuzzification method with the centreof the area (Jang et al., 1997).By dividing the region under the curve with a vertical line into two equal regions using Equation 3.12.

$$\frac{\sum_i \mu(Xi)Xi}{\sum_i \mu(Xi)} \tag{3.12}$$

Where μ(*Xi*) is the value of the membership degree at the point*Xi.*

An example is used to explain the calculation of BC risk by fuzzy rules. Suppose the value of NN output 1 (for the benign type) is "0.8" and NN output 2 (for the malignant type) is "0.2". First, we find that the linguistic variable of NN-Output 1 is *high* with respect to Fig. 3.31 and then calculate the degree of MFs using Equation 3.13.

$$\mu_{high}(0.8) = \frac{0.8-0.6}{0.3} = 2/3 \quad (3.13)$$

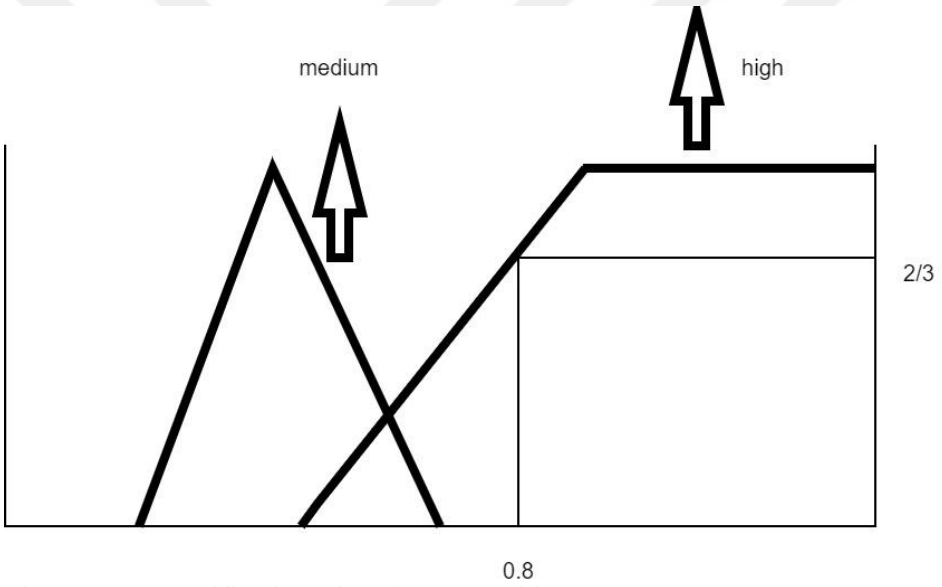


Figure 3.31. Fuzzification of benign probability value

On the other hand, the linguistic variable of NN-Output 2 is found as *low* with respect to Fig. 3.32 and the degree is computed using Equation 3. 14.

$$\mu_{low}(0.2) = \frac{0.4-0.2}{0.3} = 2/3 \quad (3.14)$$

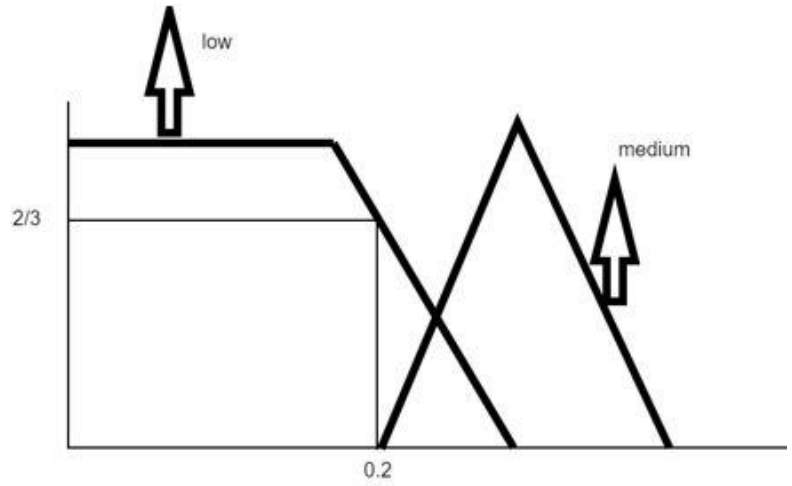


Figure 3.32. Fuzzification of malign probability value

In the next step, using the seventh rule of the algorithm, we find that the linguistic variable for risk is *very low* because the linguistic variables for benign and malignant are *high* and *low*, respectively. If you look at the 7th rule of the algorithm, you can see that it includes *AND* as an operator. So we choose the minimum of the degrees of MFs to find the degree of MF for *very low*.

$$\mu_{very\_low} = \min\left(\frac{2}{3}, \frac{2}{3}\right) = \frac{2}{3}$$

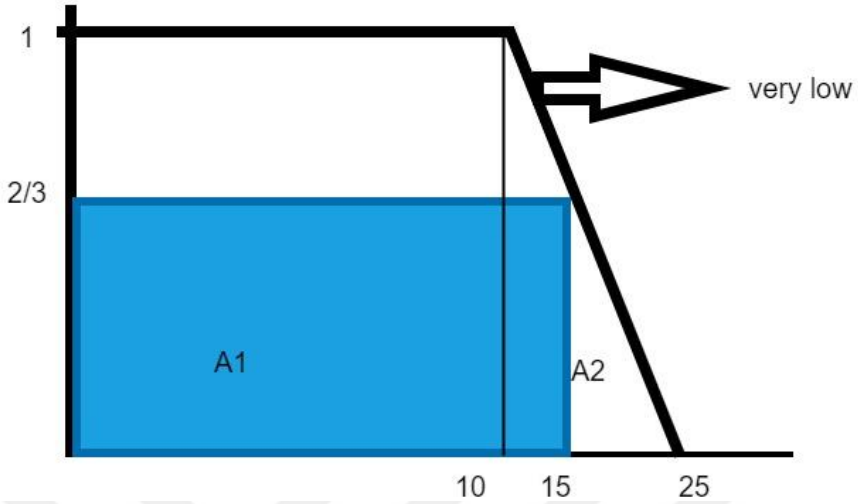


Fig.3.33. Defuzzification of the risk score

Fig. 3.33 shows the defuzzification of the linguistic variable *very low* for the risk score (RS). Using the center of area method, the risk score is determined as "10.2". This risk score is then used in a simple rule-based classifier created by a physician to determine the class of the mammogram. If the calculated risk score is greater than “50” and less than or equal to “50”, the mammograms are classified as malignant and benign, respectively, by the simple rule-based classifier.

3.2.6.3. Text+Image and Image+Image with Late Fusion and VGG16+Transformer

In this thesis, we also used the combination of text and image data with the Transformer model to determine whether the prediction of breast cancer can be improved by the late fusion method, which combines the results of each model at the classification level. With the help of radiologists, we identified the region of interest in the mammograms. We then combined the results of the mammograms and the cropped mammograms using the late fusion method. Fig. 3.34 and Fig. 3.35 show the workflow of text+image and image+image using the late-fusion

method, respectively. There is no study in the literature that combines the results of mammograms and radiological reports or cropped mammograms by late fusion for breast cancer classification.

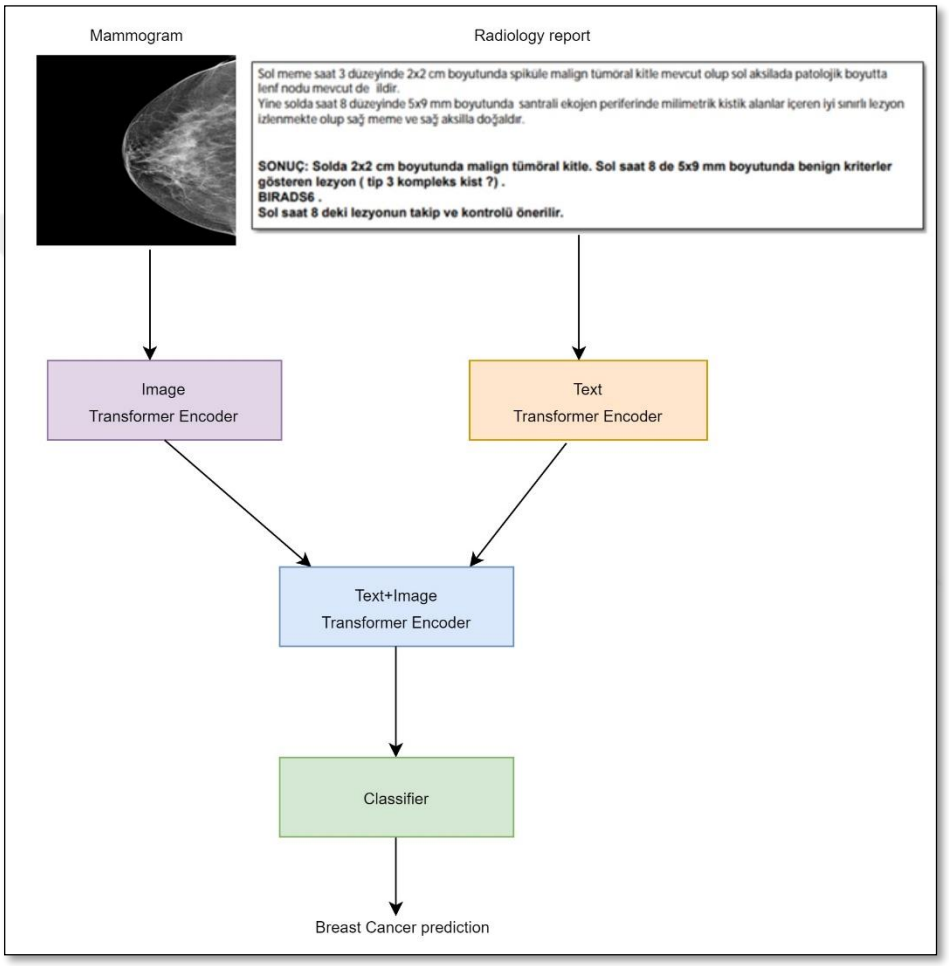


Figure 3.34.The workflow of late fusion for text+image.

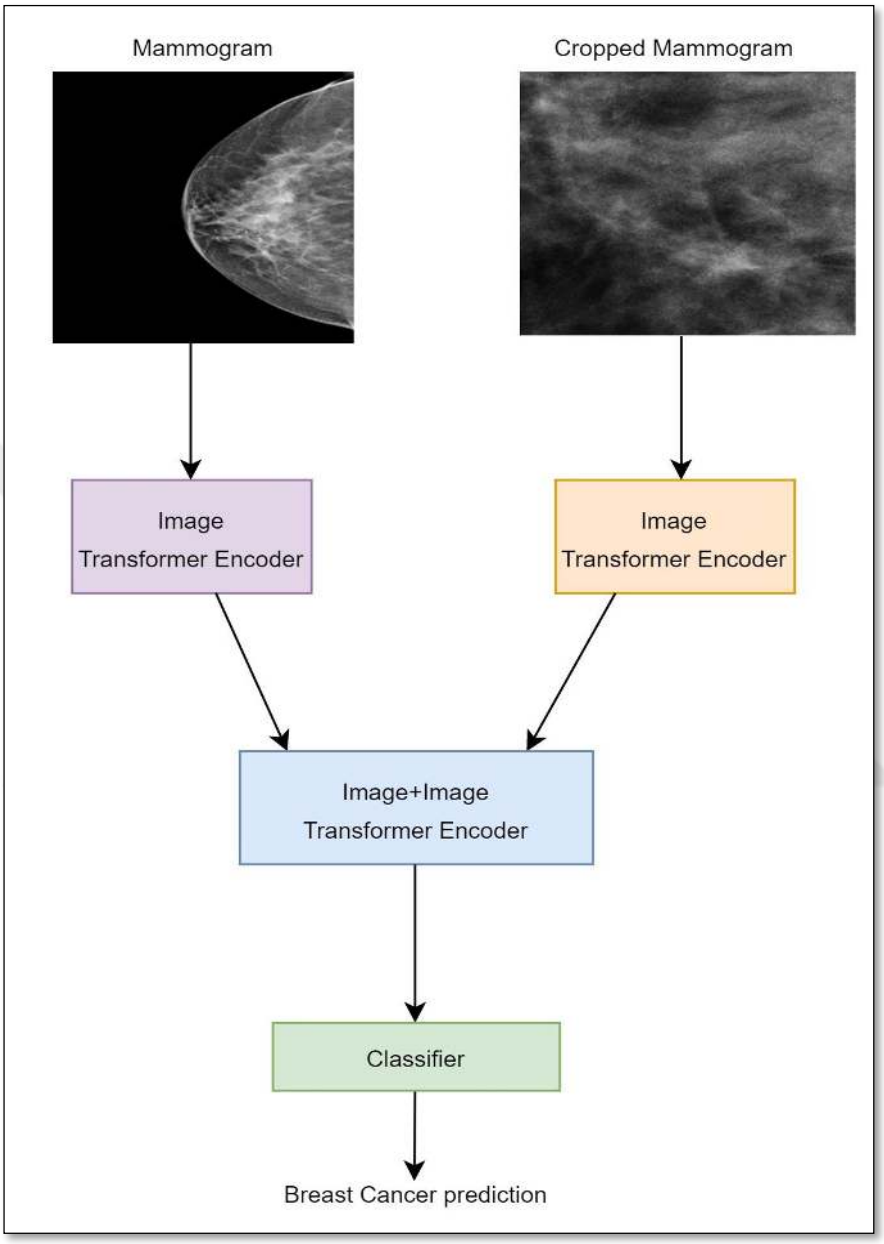


Figure 3.35. The workflow of late fusion for image+image.

In addition to analyzing various combinations such as text-image and image-image combinations by late fusion, this thesis also uses a hybrid model of pre-trained + transformer to classify mammograms as can be seen in Fig. 3.35. First, we extracted features from mammograms using the most successful pre-trained model VGG16 and then used these features as input to the transformer. Similar to the literature, we extracted features using VGG16. Unlike in the literature, we train the transformer from scratch using the extracted features. However, in the literature, researchers train vision transformer that is pretrained on image data with extracted features.

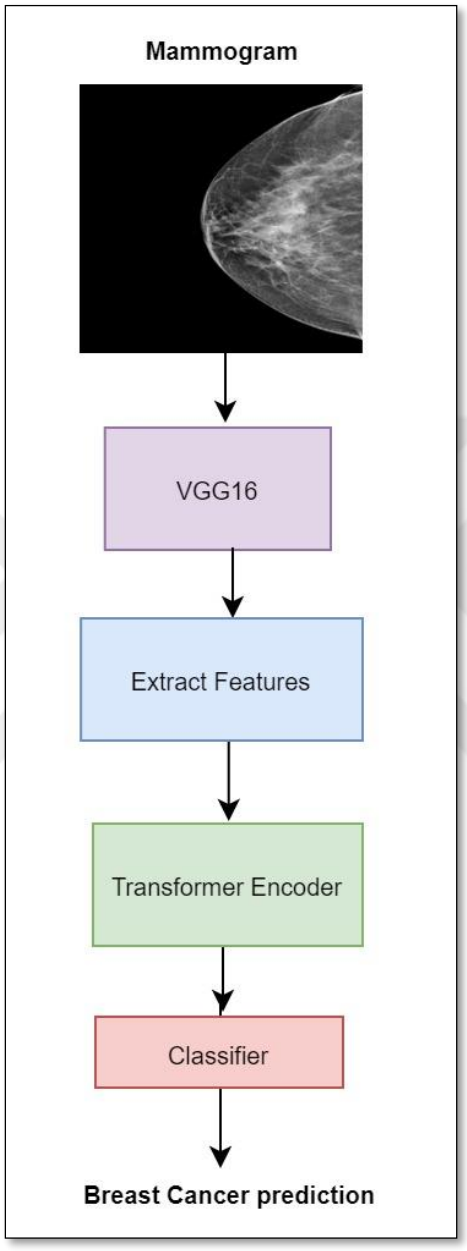


Figure 3.36. The workflow of the hybrid VGG16+Transformer

3.2.7. Hyperparameters

Hyperparameters control the learning process and determine the values of the model parameters that a learning algorithm eventually learns. Learning rate, batch size, optimization algorithms, loss functions, etc. can be an example of hyperparameters. These parameters affect the quality of learning and the speed of the learning process. In this thesis, we experimented with different types of data: Text, Image and Survey. For each type of data, different classification models and different hyperparameters are used. Table 3.16 and Table 3.17 show the hyperparameters of the pre-trained and transformer models for text and image data. All these parameters were obtained using grid search. It is used to determine the hyperparameters of the model. In grid search, the model is trained and the results are observed for the combinations of all values in the specified range.

Table 3.16. Hyperparameters used in training of pretrained models.

Hyperparameter	**Text**	**Image**	**Combined**
max length	300	-	300
Leaning rate	0.0001	0.0001	0.0001
K fold	5	5	5
Batch size	16	16	16
Epoch size	100	100	100
Patience	50	50	50
Scheduler step size	1	1	1
Gamma	0.95	0.95	0.95

Table 3.17. Hyperparameters used in training of transformer models.

Hyperparameter	**Text**	**Image**	**Combined**
max length	300	-	300
Leaning rate	0.0001	0.0001	0.0001
K fold	5	5	5
Batch size	16	16	16
Epoch size	100	100	100
Patience	50	50	50
Scheduler step size	1	1	1
Gamma	0.95	0.95	0.95
Patch size	-	14	14
D model	1	1	1
N layers	1	1	1
N hidden	1	1	1
N heads	1	7	2

4. RESULTS AND DISCUSSIONS

In this section, experimental results performed on surveys, mammogram images, and radiological reports are given in detail and the experimental results are discussed.

4.1. Experimental Results of Surveys

4.1.1. Classical Machine Learning Methods

In this section, the surveys are tested using machine learning methods. In this thesis, the augmented and normalized data are classified using traditional classifiers such as (SVM, NB, RF, etc.) in WEKA. In this study, we show the average results of 5-fold cross validation for different classifiers. All the results of the applied methods can be found in the appendix. In addition, the ensemble methods and the Attribute selected classifier give the best results among the applied classifiers.

In analyzing the overall experimental results of classifying survey data for 5-fold cross-validation, we select the method with the lowest false-negative and false-positive rates and the highest f-measure value obtained with different combinations of the different augmentation methods, normalization and feature selection techniques for 5-fold validation. In Table 4.1, we show the best experimental results by different combinations of the different augmentation methods, normalization and feature selection techniques for 5-fold validation to select the most successful combination with the classification algorithm. For example, the combination of first method augmentation+second method normalization+third method class balancing means that we first augmented the data, then normalized the augmented method, and finally balanced the augmented and normalized data. From Table 4.1, it can be seen that using pure data with Attribute selected classifier algorithm give the best result. This is because the

weighted averages of Attribute selected classifierhas the highest values for f-score, recall, precision, and accuracy, while having the lowest false negative and false positive rates. Fig. 4.1 shows the confusion matrix obtained fromAttribute selected classifier algorithm using pure data.

Table 4.1. The best 5 results of over all the experiments.

Combination*	**Algorithm**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
A	**Attributeselected**	**0.84**	**0.2**	**0.88**	**0.84**	**0.86**
B	Bagging	0.83	0.28	0.86	0.83	0.84
C	Attributeselected	0.83	0.33	0.85	0.83	0.84
D	Adaboost	0.81	0.25	0.84	0.81	0.82
E	Adaboost	0.81	0.30	0.83	0.81	0.82

*Definition of:

A. Pure data without augmentation+without Normalization
B. The first method augmentation + The second method class balancing with SMOTE + The third method normalization with Standardization
C. The first method augmentation + The second method normalization with MinMax
D. The first method class balancing with SMOTE +The second method normalization with MinMax
E. The first method augmentation + The second method class balancing with SMOTE + The third method normalization with Standardization

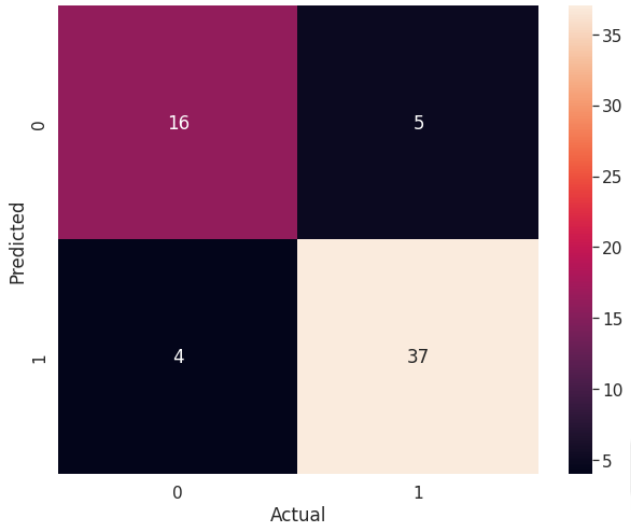


Figure 4.1. The confusion matrix of the attribute selected classifier using pure data without augmentation+without Normalization.

4.2. Experimental Results of the Radiology Reports

In order to justify the results and to discuss some observations about radiology report classification task, we present a detailed analysis in this section.

4.2.1. Classical Machine Learning Methods

As baseline results, we first apply classical machine learning models on the pure dataset. The results of the pure dataset, without applying augmentation and weighted random sampling, using machine learning algorithms are presented in Table 4.2.

As can be seen in Table 4.2, it can be said that SGD is the best algorithm, while Random Forest is the worst algorithm for the pure dataset. SVM and Logistic Regression provide the second best F1-score of 70% for the pure dataset. However, Multinomial Naive Bayes is the second worst algorithm for the pure dataset

Table 4.2. The results of classical machine learning algorithms on the pure dataset.

Weighted Average/TF-IDF	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	0.65	0.65	0.65	0.65
SVM	0.69	0.71	0.69	0.70
SGD	**0.71**	**0.72**	**0.71**	**0.71**
LogisticRegression	0.67	0.73	0.67	0.70
RandomForest	0.63	0.61	0.63	0.62

Table 4.3. The results of classical machine learning algorithms on the augmented and balanced dataset.

Weighted Average/TF-IDF	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	**0.79**	**0.80**	**0.79**	**0.79**
SVM	0.77	0.77	0.77	0.77
SGD	0.75	0.77	0.75	0.76
LogisticRegression	0.77	0.79	0.77	0.78
RandomForest	0.75	0.73	0.75	0.74

We also apply classical machine learning models on the augmented and balanced dataset by applying EDA and weighted random sampling methods. The results are given in Table 4.3. When comparing the results of the machine learning algorithms for the augmented and balanced dataset, we can see that Multinomial Naive Bayes is the best algorithm in terms of accuracy, precision, recall, and F1-score.

With Multinomial Naive Bayes we get the best precision of 80% and the best F1-score of 79%. Logistic Regression also gives the second best precision of 79% and F1-score of 78%. SVM has the second best recall that is equal to 77%, while the algorithms SGD and Random Forest have the worst recall which is equal to 75%. Random Forest achieves the worst precision rate of 73% and F1-score of 74%. From the results of the data enriched with the EDA and sampling methods, we can conclude that the use of Multinomial Naive Bayes diagnoses breast cancer with the highest accuracy. In other words, Multinomial Naive Bayes predicts most malignant findings as malignant, while Random Forest predicts most malignant findings as benign. Thus, based on these results, we can say that Multinomial

Naive Bayes is more successful in detecting BC at an early stage than Random Forest. Also, the use of Logistic Regression improves the detection of benign findings in the radiology dataset.

We also observe that when using Multinomial Naive Bayes for the radiology dataset, the number of false positives is the lowest, while it is the highest for SGD and Random Forest. These results show that most of the benign findings are predicted as benign using Multinomial Naive Bayes, while they are predicted as malignant using SGD and Random Forest. These results obtained with SGD and Random Forest may lead to unnecessary biopsy procedures. Therefore, by using Multinomial Naive Bayes for the radiology dataset, we can reduce unnecessary breast biopsies.

Comparing the results of Table 4.2 and Table 4.3, we can say that the results of the enriched dataset are better than the results of the pure dataset. Thereby, unnecessary breast biopsies can be detected more accurately with Multinomial Naive Bayes, while the use of Random Forest can lead to a lower success rate in reducing unnecessary breast biopsies. These results show that the use of augmentation and balancing the dataset have a positive impact on reducing unnecessary breast biopsies.

4.2.2. BERT models

Table 4.4 shows the results of the pure dataset with BERT models. If we compare the versions of BERT, we can see that $BERT_{Clinical}$has the worst results while $BERT_{Multilingual}$ has the best result. Moreover, $DistilBERT_{Turkish}$ has the worst results except precision score. On the other hand, Hard Votingis applied to the results of $BERT_{Multilingual,}$ BERTClinical, and BERTTurkish to improve the classification results. However, we could not get better results with Hard Voting than with $BERT_{Multilingual}$. Consequently, $BERT_{Multilingual}$ is the most successful model compared to the others. However, comparing Table 4.4 with Table 4.2, we

see that the results of the pure dataset with BERTs are better than the results of the pure dataset with machine learning algorithms.

Table 4.4.The results of BERT models and hard voting on the pure dataset.

	Accuracy	**Precision**	**Recall**	**F1-Score**
$DistilBERT_{Turkish}$	0.69	0.84	0.53	0.65
$BERT_{Turkish}$	0.77	0.73	0.74	0.73
$BERT_{Clinical}$	0.76	0.66	0.75	0.70
$BERT_{Multilingual}$	**0.94**	**0.91**	**0.94**	**0.92**
Hard Voting	0.92	0.90	0.91	0.90

Table 4.5. The results of BERT models and hard voting on the augmented and balanced dataset.

	Accuracy	**Precision**	**Recall**	**F1-Score**
$DistilBERT_{Turkish}$	0.73	0.86	0.58	0.69
$BERT_{Turkish}$	0.90	0.91	0.89	0.89
$BERT_{Clinical}$	0.87	0.87	0.87	0.87
$BERT_{Multilingual}$	**0.93**	**0.94**	**0.91**	**0.92**
Hard Voting	0.92	0.90	0.93	0.91

We also apply contextualized word embeddings (BERT versions and DistilBERT) to our augmented and balanced dataset, and the results are shown in Table 4.5. Comparing the three versions of BERT, we find that the performance of the $BERT_{Clinical}$ model is the worst, while $BERT_{Multilingual}$ has the best results in terms of accuracy, recall, precision and F1-score among the models from BERT.

According to these results, the numbers of false positives and false negatives are lowest for $BERT_{Multilingual}$, while the numbers of false negatives and false positives are highest for $BERT_{Clinical}$. The enrichment of the collected data with the EDA method may affect the Turkish words used in the augmented reports. This is because the F1-score of $BERT_{Turkish}$ is lower than $BERT_{Multilingual}$. These results show that $BERT_{Clinical}$ and $BERT_{Turkish}$ are more successful in classifying malignant reports with EDA enhancement, while $BERT_{Multilingual}$ is the most successful in classifying benign reports. Thus, we can conclude that using the EDA

approach to expand the collected Turkish dataset has an impact on the number of benign and malignant reports found correctly depending on the domain. We also test $\text{DistilBERT}_{\text{Turkish}}$ with the EDA-enriched and balanced dataset. As can be seen in Table 4.4, $\text{DistilBERT}_{\text{Turkish}}$ gives the worst results.

In addition to these BERT models, we also use the ensemble method of Hard Voting to improve breast cancer detection. However, we could not achieve better results with Hard Voting than with $\text{BERT}_{\text{Multilingual}}$. According to these classification scores, we can say that $\text{BERT}_{\text{Multilingual}}$ is the most successful one in classifying benign reports as benign and malignant reports as malignant. If we compare Table 4.4 with Table 4.5, we see that the results of the pure dataset are lower than the results of the EDA-enriched and balanced dataset. This shows that the EDA augmentation and balancing methods increase the success of the BERT models.

Also, when comparing the models of BERT with the baseline models, as can be seen in Table 4.4and Table 4.5, the accuracy, precision, recall, and F1-score values of $\text{BERT}_{\text{Multilingual}}$ are higher than the accuracy, precision, recall, and F1-score values of Multinomial Naive Bayes. Moreover, the precision score obtained with $\text{BERT}_{\text{Multilingual}}$ is higher than that obtained with the baseline models. The overall results show that using the $\text{BERT}_{\text{Multilingual}}$ model improves the classification performance of predicting the benign cases from radiological reports.

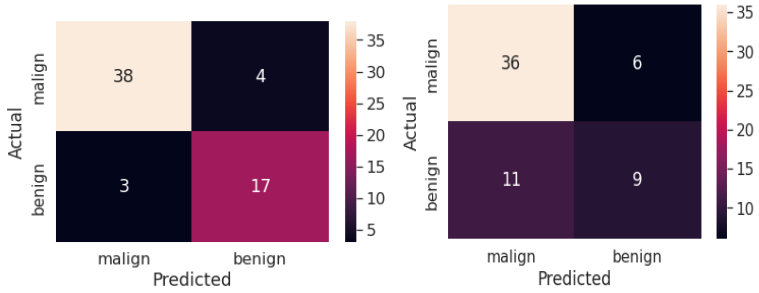
The confusion matrix for three versions of BERT, DistilBERT, and Hard Voting resulting from the EDA-enriched and balanced dataset is shown in Fig. 4.2. As you can see from the figure, the number of true positives is the lowest, and the number of false negatives is the highest for the $\text{BERT}_{\text{Clinical}}$ model. This result shows that the $\text{BERT}_{\text{Turkish}}$, $\text{BERT}_{\text{Multilingual}}$ and $\text{DistilBERT}_{\text{Turkish}}$ models classify the malignant reports more correctly. On the other hand, the number of true-negatives is the lowest in the $\text{DistilBERT}_{\text{Turkish}}$ model, while the number of false-negatives is the lowest in the $\text{BERT}_{\text{Multilingual}}$ model. Based on these results, we can say that benign reports are classified more correctly with the $\text{BERT}_{\text{Multilingual}}$ model than

with the other BERT models. In addition, we can say that the number of false-positive reports increases while the number of true-negative reports decreases when we use the Hard Voting.

When we compare the experimental results of the augmented and balanced data with the pure data, it is clearly seen that using pure data decreases the classification results of pre-trained BERT models.

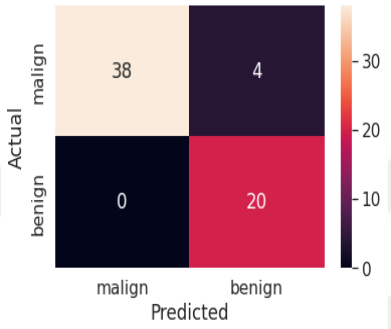
From the overall results, we can conclude that using $BERT_{Multilingual}$ on the augmented and balanced data is the most successful model in detecting breast cancer according to the interpreted reports of radiologists. In doing so, we can find that $BERT_{Multilingual}$ provides the most successful results in classifying the collected Turkish radiology reports compared to the other methods tested in this study.

Fig.4.3 shows the confusion matrices of BERT models for the pure dataset. As it can be seen from the figure, the number of true-negative reports is zero in each model, except for the $BERT_{Multilingual}$ model. The number of false-negative reports is also zero in all models.

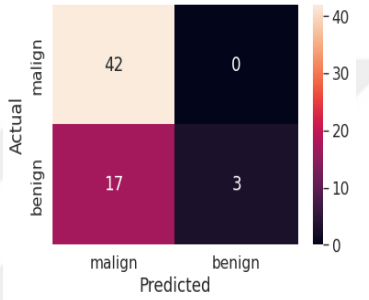


(a) Turkish BERT Model

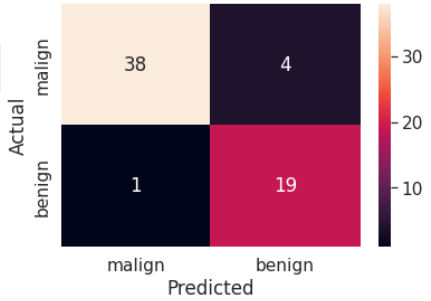
(b) Clinical BERT Model



(c) Multilingual BERT Model

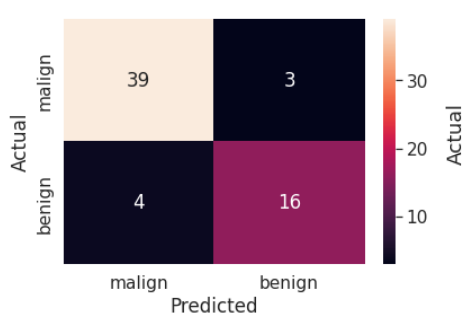


(d) Distil BERT Model

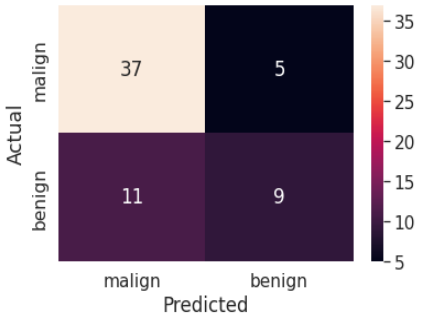


(e) Hard Voting

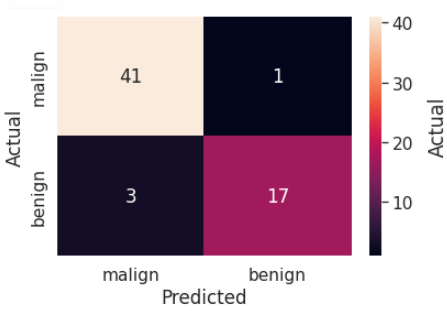
Figure4.2. Confusion matrices obtained from the augmented and balanced dataset by using (a) Turkish BERT Model (b) Clinical BERT Model (c) Multilingual BERT Model (d) Distil BERT Model and (e) Hard Voting.



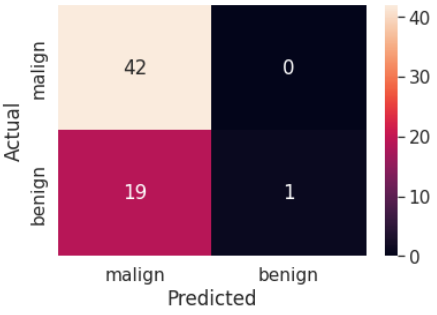
(a) Turkish BERT Model



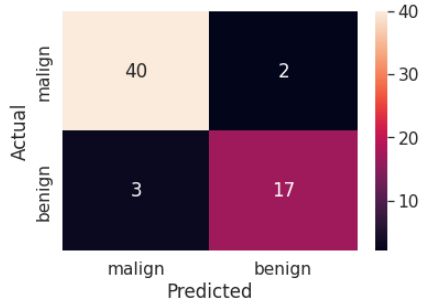
(b) Clinical BERT Model



(c) Multilingual BERT Model



(d) Distil BERT Model



(e) Hard Voting

Figure 4.3. Confusion matrices obtained from the pure data by using (a) Turkish BERT Model (b) Clinical BERT Model (c) Multilingual BERT Model (d) Distil BERT Model and (e) Hard Voting.

4.2.3. Transformer models

We investigate which pre-trained BERT model performs better as an embedding for radiology report data. The results with pre-trained embeddings for

radiology report classification using pure data transformer models are shown in Table 4.6. The results show that the best result for radiology report data is obtained with $BERT_{Multilingual}$. Transformer models with $BERT_{Clinical}$ performed the worst for this model. The reason for this performance might be related to the domain of BERT models. Domain of BERT models affects the performance. Table 4.7 shows the results of the Transformer models using augmented and balanced data. Again, the best performance was obtained with the $BERT_{Multilingual}$ model. A comparison of Table 4.6 and Table 4.7 shows that the results of the Transformer models using augmented and balanced data are better than the results of the Transformer models using pure data. The reason for this could be the training size, which is important for Transformer models.

Table 4.6. The results of Transformer models on pure dataset.

	Accuracy	**Precision**	**Recall**	**F1-Score**
Transformer_Text($BERT_{Turkish}$)	0.87	0.87	0.85	0.86
Transformer_Text($BERT_{Clinical}$)	0.81	0.78	0.78	0.78
Transformer_Text($BERT_{Multilingual}$)	**0.94**	**0.90**	**0.96**	**0.92**

Table 4.7. The results of Transformer models on augmented and balanced dataset.

	Accuracy	**Precision**	**Recall**	**F1-Score**
Transformer_Text($BERT_{Turkish}$)	0.95	0.94	0.95	0.95
Transformer_Text($BERT_{Clinical}$)	0.95	0.95	0.95	0.95
Transformer_Text($BERT_{Multilingual}$)	**0.97**	**0.97**	**0.97**	**0.97**

If we compare the experiments for the Transformer models with the BERTs, we can say that the Transformer models are more successful than the BERTs. However, when we go through the whole experimental results of radiology report classification, the Transformer model with $BERT_{Multilingual}$ using extended and balanced data shows the best performance.

4.3. Experimental Results of Mammogram Images

In this section, we present detailed experimental results on the classification of mammogram images in two separate parts, namely pre-trained models and transformers.

4.3.1. Pre-Trained Models

We classify mammograms with pre-trained models, VGG16, ResNet50, and AlexNet. Table 4.8 shows the experimental results for the pre-trained models without weighted random sampling method used to balance the data. The performance of the pre-trained models without weighted sampling method with/without augmentation by simple transformations is shown in Table 4.8. Table 4.8 shows that AlexNet without augmentation has the highest F1 score of 74%, while AlexNet with augmentation has the lowest F1 score of 59%. This shows that augmenting the training data without balancing affects the performance of the model. VGG without augmentation shows the second best F1 score of 73% and precision of 72%, while VGG with augmentation gives the highest recall of 80% and accuracy of 79%. Thus, the AlexNet model with pure dataset (training data without augmentation and without weighted random sampling) shows the highest values for classification of mammograms.

In this section, we present detailed experimental results on the classification of mammogram images in two separate parts, namely pre-trained models and transformers.

Table 4.8. The results of Pre-Trained models without weighted random sampling.

Models	Augmentation	Accuracy	Precision	Recall	F1-Score
VGG	No	0.77	0.72	0.75	0.73
VGG	Yes	0.79	0.70	0.80	0.75
ResNet	No	0.77	0.69	0.77	0.73
ResNet	Yes	0.76	0.69	0.73	0.71
AlexNet	**No**	**0.79**	**0.72**	**0.77**	**0.74**
AlexNet	Yes	0.68	0.58	0.61	0.59

Table 4.9. The results of Pre-Trained models with weighted random sampling.

Model	Augmentation	Accuracy	Precision	Recall	F1-Score
VGG	No	0.97	0.97	0.97	0.97
VGG	**Yes**	**0.98**	**0.98**	**0.98**	**0.98**
ResNet	No	0.89	0.88	0.89	0.88
ResNet	Yes	0.90	0.90	0.90	0.90
AlexNet	No	0.97	0.97	0.96	0.96
AlexNet	Yes	0.90	0.90	0.90	0.90

On the other hand, Table 4.9 shows the results of the pre-trained models with/without augmentation for weighted random sampling. As shown in Table 4.9, VGG with augmentation has the highest values accuracy of 97%, precision, recall, and F1-score of 98%. However, ResNet without augmentation has the lowest values for accuracy, precision, recall, and F1-score. The results of AlexNet without augmentation are higher than AlexNet with augmentation. Nevertheless, the results of VGG and ResNet without augmentation are lower than the results of VGG and ResNet with augmentation. Comparison of Table 4.8 and Table 4.9 shows that the results of pre-trained models using weighted random sampling are higher than the results of pre-trained models without weighted random sampling. As a result, VGG using augmentation and weighted random sampling shows the best result for mammogram image classification overall pre-trained models. Fig. 4.4 shows the confusion matrix of the best model.

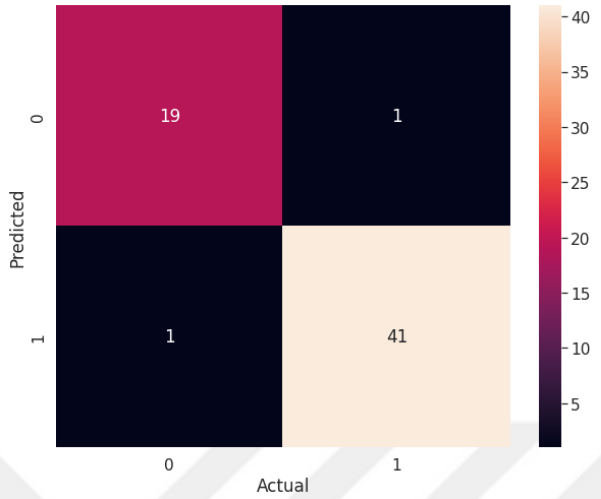


Figure 4.4. Confusion matrix of the VGG with augmentation with weighted random sampling.

4.3.2. Transformer Models

A comparison of the transformer model with/without augmentation and with/without weighted random sampling is shown in Table 4.10. This table shows that the transformer model with pure data (without augmentation and without weighted random sample) gives the best result. We achieved 79% of accuracy, precision, recall, and F1 score. When we use the weighted random sampling method for the training data, we again achieve higher results with the Transformer model without augmentation of the training data. When we compare the results without weighted random sampling and with weighted random sampling, the results of Transformer with balancing training data with weighted random sampling are lower than the results of Transformer without weighted random sampling. This shows that the transformer model without weighted random sampling learns better.

Table 4.10. The results of Transformer modelwith/without augmentation and with/without weighted sampling.

Weighted Random Sampling	Augmentation	Accuracy	Precision	Recall	F1-Score
No	No	**0.79**	**0.79**	**0.79**	**0.79**
No	Yes	0.77	0.75	0.74	0.75
Yes	No	0.71	0.59	0.67	0.63
Yes	Yes	0.69	0.56	0.64	0.60

4.3.2.1. Experimental Results of Cropped Mammogram Images

In this section, we analyze the cropped image data using the transformer. We test the transformer with and without applying weighted random sampling and augmentation for cropped image data. As can be seen in Table 4.11, we get higher values with weighted random sampling than without weighted random sampling. With weighted random sampling and augmentation, we get the highest score, while with the augmentation and without weighted random sampling we get the lowest score. We obtain values of 80%, 75%, 77%, and 76% for accuracy, precision, recall, and F1 score, respectively, when we use the transformer with weighted random selection and augmentation for the cropped image data. Comparing the results of the transformer for mammogram images and the results of the transformer for cropped mammogram images, we can see that the transformer with mammogram images is more successful than the transformer with cropped mammogram images.

Table 4.11. The results of Transformer Image (Cropped).

Weighted Random Sampling	Augmentation	Accuracy	Precision	Recall	F1-Score
No	No	0.72	0.50	0.44	0.47
No	Yes	0.69	0.55	0.86	0.67
Yes	No	0.72	0.68	0.68	0.68
Yes	Yes	**0.80**	**0.75**	**0.77**	**0.76**

We also test the performance of concatenating cropped images and image data using the fusion method. As can be seen in Table 4.12, the Transformer model without augmentations performs better than the Transformer model with the augmentations, both with and without applying weighted random selection. Using the Transformer model for cropped images and image combinations, the highest F1 score of 66% is obtained without augmenting the training data and with applying weighted random selection. The comparison of Table 4.11 and Table 4.12 shows that the results of the transformer model with cropped data are higher than the results of the transformer model with the combination of cropped images and images with the fusion method. In conclusion, using the fusion method combining cropped images and image data has a negative impact on the classification of mammogram images.

Table 4.12. The results of Transformer model the combination of Image and Image (Cropped).

Weighted Random Sampling	**Augmentation**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
No	No	0.70	0.58	0.64	0.61
No	Yes	0.60	0.51	0.51	0.51
Yes	No	**0.68**	**0.67**	**0.65**	**0.66**
Yes	Yes	0.68	0.67	0.65	0.66

4.3.2.2. Experimental Results of Merged Mammogram Images

Cropped images of hospital data, MIAS data, and BCDR-D02 data are merged to have larger training data size. Because the size of training data is important in transformer model. Table 4.13 shows the results of transformer model using merged data. As shown in Table 4.13, we have the highest scores such as accuracy, precision, recall, and F1-score of 78%. When we compare Table 4.12, and Table 4.13, we have higher results merging cropped mammogram images than

using only cropped mammogram image. This result shows that transformer model performs better when using larger training size.

Table 4.13. The results of Transformer model using the merged Mammogram images.

Weighted Random Sampling	**Augmentation**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
No	No	0.74	0.73	0.73	0.73
No	Yes	0.75	0.74	0.74	0.74
Yes	No	0.75	0.75	0.75	0.75
Yes	Yes	**0.78**	**0.78**	**0.78**	**0.78**

4.4. Experimental Results of Combination of Radiology Reports and Mammogram Images

In this part, we test the performance of concatenating text and image with fusion. Table 4.14 shows the results with and without weighted random sampling. The transformer with the combination of text and image model without augmentation shows the best values accuracy of 96%, precision of 97%, recall of 94% and F1 score of 95% with weighted random sampling, while the model with augmentation shows the lowest values accuracy of 84%, precision of 79%, recall of 83% and F1 score of 81% without weighted random sampling. However, the model without augmentation performs better than the model with augmentation both with and without weighted random sampling. In addition, the model with weighted random sampling shows higher scores than the model without weighted random sampling. As a result, we get the best results for the transformer model with a combination of text and image using weighted random sampling without augmentation of the training data.

Table 4.14. The results of Transformer model the combination of Text and Image.

Weighted Random Sampling	Augmentation	Accuracy	Precision	Recall	F1-Score
No	No	0.85	0.84	0.83	0.83
No	Yes	0.84	0.79	0.83	0.81
Yes	No	**0.96**	**0.97**	**0.94**	**0.95**
Yes	Yes	0.90	0.90	0.90	0.90

4.5. Experimental Results of Hybrid of All Three Data Types

In this part, we test the performance of the developed hybrid model for breast cancer classification. We used each patient's mammogram, report, and survey as input to the hybrid model. This hybrid model uses the results of the most successful model as input for three types of data: Image, Text, and Survey. As shown in Table 4.15, we use the results of the pre-trained model for mammograms, the results of the transformers for reports, and the results of the pure data with Attribute selected classifier for surveys as input. The results of the hybrid model show very high values in predicting malignant patients compared to the best models using only survey, report or image data to reduce the cost of breast biopsies. The hybrid model with the hospital dataset improves the precision value by up to 100%. This shows that our hybrid model has the fewest false-positive instances for our dataset. In other words, the hybrid model helps reduce the number of patients who do not have breast cancer (benign) but are classified as breast cancer (malignant). In addition, we achieved a higher accuracy of 98% and an F1 score of 98% with the hybrid model than with the Transformer models. However, the precision of the hybrid model is higher than that of the pre-trained models, while the accuracy, recall, and F1 score of the hybrid model are identical to the results of the pre-trained models. On the other hand, the accuracy, precision, recall, and F1 score of the hybrid model are higher than the results of the survey-only data with Attribute selected classifier. Table 4.16 shows the actual label and the predicted label of the test patients. We take the results of the best models for each data type (text, image, and survey) and generate an output label for each patient

using ensemble learning, as shown in Table 4.16. Example: test patient 12: While the predicted label of ImageVGG_Pretrained is malignant, the predicted label of TextTransformer and SurveyPureData_Attributeselected is benign, so the predicted label of the hybrid model is benign. Fig. 4.5 shows the confusion matrix of the hybrid model. In this figure, 0 represents the benign class type and 1 represents the malignant class type.

Table 4.15. The results of hybrid model.

	Accuracy	**Precision**	**Recall**	**F1-Score**
($Image_{VGG\_Pretrained}$+$Text_{Transformer}$+$Survey_{Pure\ Data\_Attributeselected}$)	**0.98**	**1**	**0.98**	**0.98**

Table 4.16. The actual label and the predicted label according to the hybrid model for one fold.

Test instances	Actual label	Predicted label
1	malign	malign
2	malign	malign
3	malign	malign
4	malign	malign
5	benign	benign
6	benign	benign
7	benign	benign
8	benign	benign
9	malign	malign
10	malign	malign
11	malign	malign
12	malign	benign +
13	malign	malign

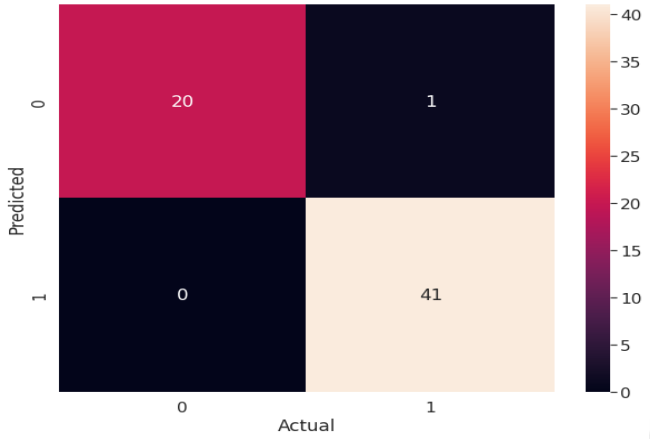


Figure 4.5. Confusion matrix obtained from the hybrid model.

4.6. Experimental Results of Risk-based Hybrid Neuro+FuzzyRule-based Model

A CNN (as a base) and ResNet50 neural networks are used in the DL part of the hybrid system to show how successful the proposed risk-based model is in detecting BC. For the deep learning part, we chose ResNet50 instead of VGG16. Since VGG16 is the most successful model in classifying mammogram images, we have shown how the fuzzy approach increases the precision values of the worst model, ResNet50, since VGG16 will achieve higher results than ResNet50 in this proposed model with fuzzy approach. These neural networks are trained to find the inputs to the second phase (FRBS) of the hybrid model. FRBS finds the risk values to detect ambiguous data. When we compare the experiments of the proposed model, we use CNN and ResNet50 as the base models.

The experiments with the proposed hybrid model and the base models (CNN and Resnet50) are performed on the BCDR-D02 and merged datasets. We use simple transformations to increase the sample size of the data. Since the number of malignant type samples is not equal to the number of benign type samples, the under sampling method is used for data balancing. For each test sample, a class label is assigned based on the calculated risk values and a simple

rule-based classifier. The scores of evaluation metrics are then calculated and compared to the basic neural network evaluation metrics.

The randomly selected NN-Outputs 1 and 2 and the calculated risk values of the proposed risk-based hybrid system corresponding to the NN outputs are shown in Table 4.17. These NN outputs were determined using the CNN. As you can see in Table 4.17, the percentage of risk is higher in the cases where the probability value of the malignant type is higher than that of the benign type. If we look at condition 1 in Table 4.17, we see that the percentage of risk is 76% when the outputs for the benign type and the malignant type are 0.36 and 0.89, respectively. After determining the percentage of risk, we use this percentage in a simple rule-based classifier that determines the class of states. According to our proposed model, condition 6 is classified as malignant because the percentage of risk is 79%, which is higher than 50%.

Table 4.17. Risk results of the breast cancer using CNN.

Conditions	**Benign**	**Malignant**	**Risk(%)**
Condition1	0.36	0.89	76
Condition2	0.52	0.35	47
Condition3	0.74	0.63	34
Condition4	0.48	0.41	50
Condition5	0.92	0.71	44
Condition6	0.32	0.95	79
Condition7	0.87	0.69	41

Experiments are conducted with the proposed risk-based model and neural network models (CNN and ResNet50), and the evaluation metrics of these models are calculated and compared. The confusion matrices and the scores of the evaluation metrics for BCDR-D02 are shown in Table 4.18 and Table 4.19, respectively. Moreover, important results are highlighted in bold.

Table 4.18. Confusion matrices of CNN, ResNet50 and risk score based hybrid neuro+fuzzy model for BCDR-D02.

Actual Label	Predicted Label							
	CNN-only		**CNN+ Fuzzy**		**ResNet50-only**		**ResNet50+Fuzzy**	
	YES	NO	YES	NO	YES	NO	YES	NO
YES	30	12	32	10	**37**	**5**	37	5
NO	11	31	7	35	6	36	**4**	**38**

Table 4.19. Results of CNN, ResNet50 and risk score based hybrid neuro+fuzzy model for BCDR-D02.

	CNN-only	**CNN +Fuzzy**	**ResNet50-only**	**ResNet50+Fuzzy**
Accuracy	0.73	**0.80**	0.87	**0.89**
Recall	0.71	**0.76**	0.88	0.88
Precision	0.73	**0.82**	0.86	**0.90**
F1-Score	0.72	**0.79**	0.87	**0.88**

Table 4.18 shows the experimental results for CNN-only, ResNet50-only, CNN with fuzzy, and ResNet50 with fuzzy. The results of CNN-only for accuracy, recall, precision and F1-score: 73%, 71%, 73% and 72%, respectively, and according to these results we can say that CNN-only has the lowest values. Using the same CNN with a combination of FRBS (CNN+Fuzzy), we obtain values for accuracy, recall, precision, and F1-score of 80%, 76%, 82%, and 79%, respectively. As you can see from the experimental results, the scores are improved with the proposed risk-based hybrid model compared to a pure CNN.On the other hand, comparing the experiments of CNN-only and ResNet50-only, the experimental results of ResNet50-only are better than those of CNN-only. Therefore, we can conclude that the use of the transfer learning model helped to improve the results of the evaluation metrics.When we compare the results of our proposed risk-based hybrid model with the CNN-only and ResNet50 models, we find that the values for accuracy, recall, precision, and F1-score increase. We obtain the best results with a combination of ResNet50 and FRBS. With this model, we achieve a precision of 90%. Thus, we can conclude that our proposed risk-based model is important to prevent biopsies that are not malignant. And while

our model achieves an F1-score of 88%, the F1-scores of the CNN-only model and the CNN+fuzzy model are 72% and 79%, respectively.

Table 4.20. Confusion matrices of CNN, ResNet50 and risk score based hybrid neuro+fuzzy model for the Merged dataset.

Actual Label	**Predicted Label**							
	CNN-only		**CNN+ Fuzzy**		**ResNet50-only**		**ResNet50+Fuzzy**	
	YES	NO	YES	NO	YES	NO	YES	NO
YES	84	42	92	34	**114**	**12**	115	11
NO	49	77	34	96	17	109	**13**	**113**

Table 4.21. Results of CNN, ResNet50 and risk score based hybrid neuro+fuzzy model for the Merged dataset.

	CNN-only	**CNN +Fuzzy**	**ResNet50-only**	**ResNet50+Fuzzy**
Accuracy	0.64	**0.73**	0.88	**0.90**
Recall	0.63	**0.73**	0.90	**0.91**
Precision	0.67	**0.73**	0.87	**0.90**
F-Score	0.65	**0.73**	0.88	**0.90**

The confusion matrix and experimental results for the fused dataset of the proposed risk-based hybrid model (i.e., CNN+Fuzzy and ResNet50+Fuzzy) and the classical CNN-only and ResNet50-only are shown in Table 4.20 and Table 4.21, respectively. Similar to the results on BCDR-D02, the results of the proposed risk-based hybrid model are more successful than those of CNN-only and ResNet50-only in terms of accuracy, precision, recall, and F1-score. And again, we can say that the results of the transfer learning model ResNet50 are better than the results of the classical CNN and that the best results for the merged dataset are obtained by combining ResNet50 with FRBS. This demonstrates the success of the proposed risk-based hybrid model in detecting benign data and avoiding unnecessary biopsies. The use of a larger dataset increases the score of the experimental results of the pre-trained model.

Comparing the experiments in Table 4.19 with Table 4.21, we find that the proposed risk-based system is superior to the CNN-only and ResNet50 models,

suggesting that the use of FRBS improves classification results in BC diagnosis and contributes to a reduction in unnecessary biopsies.

4.7. Experimental Results of Hybrid of VGG16 +Transformer

In this part, we test the performance of hybrid of VGG16+Transformer. Table 4.22 shows the results with and without weighted random sampling. The hybrid model with augmentationshows the best values accuracy of 76%, precision of 76%, recall of 76% and F1 score of 76% with weighted random sampling, while the model without augmentation shows the lowest values accuracy of 74%, precision of 73%, recall of 73% and F1 score of 73% without weighted random sampling. However, the model with both augmentation and weighted random sampling performs better than the model without augmentation and without weighted random sampling. And the model with weighted random sampling also shows higher scores than the model without weighted random sampling.

Table 4.22. The results of hybrid of VGG16+Transformer.

Weighted Random Sampling	**Augmentation**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
No	No	0.74	0.73	0.73	0.73
No	Yes	0.75	0.74	0.74	0.74
Yes	No	0.75	0.75	0.75	0.75
Yes	**Yes**	**0.76**	**0.76**	**0.76**	**0.76**

5. CONCLUSIONS

In this thesis, we proposed to develop a novel hbyrid method for breast cancer prediction to reduce the cost of breast biopsies by considering deep learning models such as Transformers, BERTs, VGG16, etc. One of the innovations of this thesis is that the proposed model combines all three types of data for each patient to reduce the cost of breast biopsies. However, in the literature, only one type of data set has usually been used to classify breast cancer, e.g., only image data, only radiology reports, or only survey data.

Another novelty of this thesis is that we collected a new Turkish dataset which consists of Turkish radiology reports, Turkish surveys and mammograms to test the performance of the proposed models. This new Turkish dataset collected from breast cancer patients to support breast cancer research. Since the number of radiology reports, mammogram images, and surveys is small and unbalanced, EDA expansion and self-copy for radiology reports, simple transformations for mammogram images and self-copy for surveys, and weighted random sampling techniques are used. Classification performance is evaluated experimentally for each data type separately. Traditional machine learning models, BERT models and transformer models are used to test our newly created Turkish radiology report dataset. VGG16, ResNet50 and AlexNet pre-trained models and transformer models are used to classify our newly created mammography dataset. Various machine learning algorithms are used to classify our newly created dataset of Turkish surveys. We selected the best models whose performance was highest for each data type and used the results of these models as input to the hybrid model. Our hybrid model uses the three data types (text, image, and survey) as input and produces a result of whether the patient is malignant or benign to reduce the cost of breast biopsies. In addition, we used the combination of text and image data using Transformer modelto determine whether breast cancer prediction can be improved by the late fusion method.

The experimental results show that the pre-trained VGG16 model performs the best for mammograms, the Transformer model with $BERT_{multilingual}$ performs the best for radiology reports, and Attribute selected classifier with pure data performs the best for surveys. The experimental results also show that the hybrid model performs better in predicting benign data than all other models for classifying malignant data. In other words, the proposed model achieves a precision value by up to 100% is obtained with. Thus, our hybrid model reduces the cost of breast biopsies. Therefore, we can say that our proposed model is the best choice for breast cancer detection using Turkish radiology reports, Turkish surveys, and newly collected mammograms

Another innovation of this thesis is the proposed risk-based hybrid neuro+fuzzy rule-based system that use a fuzzy approach and a rule-based system for breast cancer classification. The use of the fuzzy system has the undeniable benefit that ambiguities can be successfully analyzed and defined as real numbers. The proposed FRBS helps to reduce unneeeded biopsies by considering the probabilities generated by neural networks. To compare the classification results of the proposed system, both the merged dataset consisting of BCDR-D02, hospital data and MINI-MIAS, and only BCDR-D02 are used.

Our main objective for this proposed model is to show the success of the proposed risk-based hybrid of neural networks, fuzzy and rule-based classifiers over the neural network models used as the basis. In addition to CNN, we also used the transfer learning model ResNet50 to obtain baseline results for the proposed model. Because our data set is too small to train CNN.

We begin our experiments by testing pure CNN and pure ResNet50 models with the fused and BCDR-D02 datasets. From these experiments, we conclude that pure ResNet50 performs better than pure CNN. After testing the baseline models, we continue the experiments by testing the proposed risk-based hybrid model with both CNN and ResNet50 generated probabilities. Finally, the results of the proposed risk-based model with both CNN- and ResNet50-generated probabilities

are compared with the results of the baseline models. Based on the experimental results, we find that the proposed risk-based hybrid model with ResNet50-generated probabilities has the best precision value and f-score value.

As a result, a combination of DNN with FRBS leads to better classification results for BC diagnosis than a pure CNN and a pure ResNet50. Finally, we believe that this hybrid model will help reduce biopsy costs and the associated negative consequences such as health care costs, pain, anxiety, and so on.

We also believe that our thesis will serve as a guide that will provideto researchers in the field of breast cancer detection with useful tips for selecting the best models for classifying mammogram images, surveys, and radiology reportsin order to develop more effective and reliable studies. In the future, we plan to evaluate the performance of the proposed models with different learning models. We will analyze the proposed risk-based model using VGG16 on the hospital dataset. We also plan to use Vision Transformer to improve the results of mammogram images.

REFERENCES

Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., & Abdel-Mottaleb, M., (2021). Convolutional neural networks for breast cancer detection in mammography: A survey. Computers in biology and medicine, 131, 104248.

Agarwal, S., (2013). Data mining: Data mining concepts and techniques. In 2013 international conference on machine intelligence and research advancement (pp. 203-207). IEEE.

Ahmed, F., Adnan, M., Malik, A., Tariq, S., Kamal, F., & Ijaz, B., (2021). Perception of breast cancer risk factors: Dysregulation of TGF-β/miRNA axis in Pakistani females. *Plos one*, *16*(7), e0255243.

Alanazi, S. A., Kamruzzaman, M. M., Islam Sarker, M. N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M. H. (2021). Boosting breast cancer detection using convolutional neural network. *Journal of Healthcare Engineering*, *2021*.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M., (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Alzubaidi, L., Al-Shamma, O., M. Fadhel, A., Farhan, L., Zhang, J, and Duan, Y., 2020. "Optimizing the Performance of Breast Cancer Classification by Employing the Same Domain Transfer Learning from Hybrid Deep Convolutional Neural Network Model," Electronics, vol. 9, no. 3, Art. no. 445, DOI: 10.3390/electronics9030445.

Ao, Y., Li, H., Zhu, L., Ali, S., Yang, Z., (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. J. Petrol. Sci. Eng. 174, 776-789.

Arafat, H. M., Omar, J., Muhamad, R., Al-Astani, T. A. D., Shafii, N., Al Laham, N. A., &Jebril, M. A. A. R., (2021). Breast Cancer Risk from Modifiable

and Non-Modifiable Risk Factors among Palestinian Women: A Systematic Review and Meta-Analysis. Asian Pacific journal of cancer prevention: APJCP, 22(7), 1987.

Arifoğlu, D., Deniz, O., Aleçakır, K., &Yöndem, M., (2014). CodeMagic: semi-automatic assignment of ICD-10-AM codes to patient records. In *Information sciences and systems 2014* (pp. 259-268). Springer, Cham.

Arora, M., Tagra, D. (2012). Neuro-fuzzy expert system for breast cancer diagnosis, In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 979-985. Doi: 10.1145/2345396.2345554

Ataman, F., &Özgüner, Ö., (2021). Determination of Social Media Users' Perceptions About "Black Friday" Activity via Sentiment Analysis. *Educational Research (IJMCER)*, *3*(3), 361-371.

Baharuddin, W.N.A., Hussain, R.I., Abdullah, S.N.H.S., Fitri, N., Abdullah, A. (2013). Mamdani-fuzzy expert system for BIRADS breast cancer determination based on mammogram images, Communications in Computer and Information Science, vol. 378, pp. 99-110. Doi: 10.1007/978-3-642-40567-9_9

Balanică, V., Dumitrache, I., Caramihai, M., Rae, W., Herbst, C. (2011). Evaluation of breast cancer risk by using fuzzy logic, University Politehnica of Bucharest Scientific Bulletin, Series C, vol. 73, no. 1, pp. 53-64.

Batista, G. E., Prati, R. C., &Monard, M. C., (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

Bayrak, E. A., Kırcı, P., &Ensari, T., (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific meeting on electrical-electronics& biomedical engineering and computer science (EBBT)* (pp. 1-3). IEEE.

Bell, D.J., (December 7, 2020). American College of Radiology. *Radiopedia*.https://radiopaedia.org/articles/american-college-of-radiology?lang=us

Boroumandzadeh, M., &Parvinnia, E., (2021). Automated classification of BI-RADS in textual mammography reports. Turkish Journal of Electrical Engineering and Computer Sciences, 29(2), 632-647.

Breast cancer. (2021, March 26). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Cancer., (2022, February 3). World Health Organization.https://www.who.int/news-room/fact-sheets/detail/cancer.

Casey, A., Davidson, E., Poon, M., Dong, H., Duma, D., Grivas, A., ... & Alex, B., (2021). A systematic review of natural language processing applied to radiology reports. *BMC medical informatics and decision making*, *21*(1), 1-18.

Castro, S. M., Tseytlin, E., Medvedeva, O., Mitchell, K., Visweswaran, S., Bekhuis, T., & Jacobson, R. S., (2017). Automated annotation and classification of BI-RADS assessment from radiology reports. *Journal of biomedical informatics*, *69*, 177-187.

Chang, J., Yu, J., Han, T., Chang, H. J., & Park, E., (2017, October). A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer. In *2017 IEEE 19th international conference on e-health networking, applications and services (Healthcom)* (pp. 1-4). IEEE.

Chen, X., Zhang, K., Abdoli, N., Gilley, P. W., Wang, X., Liu, H., ... &Qiu, Y., (2022). Transformers Improve Breast Cancer Diagnosis from Unregistered Multi-View Mammograms. *Diagnostics*, *12*(7), 1549.

Choi, J. P., Han, T. H., & Park, R. W., (2009). A hybrid Bayesian network model for predicting breast cancer prognosis. Journal of Korean Society of Medical Informatics, 15(1), 49-57.

Chougrad, H., Zouaki, H., &Alheyane, O., (2017). Convolutional neural networks for breast cancer screening: transfer learning with exponential decay. *arXiv preprint arXiv:1711.10752*.

Christobel, A., &Sivaprakasam, Y., (2011). An empirical comparison of data mining classification methods. *International Journal of Computer Information Systems*, *3*(2), 24-28.

Çelıkten, A., &Bulut, H., (2021, June). Turkish Medical Text Classification Using BERT. In *2021 29th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Çetinoğlu, Ö., Bilgin, O., &Oflazer, K., (2018). Turkish Wordnet. In *Turkish Natural Language Processing* (pp. 317-336). Springer, Cham.

Decision Stump., (2021).*Wikipedia*. https://en.wikipedia.org/wiki/Decision_stump

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... &Houlsby, N., (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dromain, C., Boyer, B., Ferre, R., Canale, S., Delaloge, S., &Balleyguier, C., (2013). Computed-aided diagnosis (CAD) in the detection of breast cancer. *European journal of radiology*, *82*(3), 417-423.

Fakoor, R., Ladhak, F., Nazi, A., & Huber, M., (2013, June). Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning* (Vol. 28, pp. 3937-3949). ACM, New York, USA.

Falconi, L. G., Perez, M., Aguilar, W. G., &Conci, A., (2020). Transfer learning and fine tuning in breast mammogram abnormalities classification on CBIS-DDSM database. *Adv. Sci. Technol. Eng. Syst. J*, *5*, 154-165.

Fatima, N., Liu, L., Hong, S., & Ahmed, H., (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360-150376.

Goreke, V., Uzunhisarcikli, E., &Oztoprak, B., (2016, October). Type-2 fuzzy inference system design for computer aided detection in mammogram image. In *2016 Medical Technologies National Congress (TIPTEKNO)* (pp. 1-5). IEEE.

Grancharova, M., & Dalianis, H., (2021). Applying and sharing pre-trained BERT-models for named entity recognition and classification in Swedish electronic patient records. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 231-239).

Gupta, M., Wu, H., Arora, S., Gupta, A., Chaudhary, G., & Hua, Q., (2021). Gene Mutation Classification through Text Evidence Facilitating Cancer Tumour Detection. *Journal of Healthcare Engineering*, *2021*.

Hall, M. A., & Smith, L. A., (1998). Practical feature subset selection for machine learning.

Hall, M. A., (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

Han, J., Pei, J., & Tong, H., (2022). Data mining: concepts and techniques. Morgan kaufmann.

Hassan, S. A. A., Sayed, M. S., Abdalla, M. I., &Rashwan, M. A., (2020). Breast cancer masses classification using deep convolutional neural networks and transfer learning. *Multimedia Tools and Applications*, *79*(41), 30735-30768.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hepsağ, P. U., Özel, S. A., &Yazıcı, A., (2017, October). Using deep learning for mammography classification. In *2017 International Conference on Computer Science and Engineering (UBMK)* (pp. 418-423). IEEE.

Irvin, V. L., Zhang, Z., Simon, M. S., Chlebowski, R. T., Luoh, S. W., Shadyab, A. H., ... & Jindal, S. (2020). Comparison of mortality among participants of women's health initiative trials with screening-detected breast cancers vs interval breast cancers. *JAMA network open*, *3*(6), e207227-e207227.

Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M., & Kabir, M. N., (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*(5), 1-14.

Israni, P., (2019). Breast cancer diagnosis (BCD) model using machine learning. *Int. J. Innov. Technol. Exploring Eng.*, *8*(10), 4456-4463.

Jadoon, M. M., Zhang, Q., Haq, I. U., Butt, S., &Jadoon, A., (2017). Three-class mammogram classification based on descriptive CNN features. BioMed research international, 2017.

Jaiswal, A., Tang, L., Ghosh, M., Rousseau, J. F., Peng, Y., & Ding, Y. (2021, November). RadBERT-CL: Factually-Aware Contrastive Learning For Radiology Report Classification. In *Machine Learning for Health* (pp. 196-208). PMLR.

Jang, J. S. R., Sun, C. T., &Mizutani, E. (1997). Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]. IEEE Transactions on automatic control, 42(10), 1482-1484.

Kapoor, P. M., Lindström, S., Behrens, S., Wang, X., Michailidou, K., Bolla, M. K., ... & Breast Cancer Association Consortium, (2020). Assessment of interactions between 205 breast cancer susceptibility loci and 13 established risk factors in relation to breast cancer risk, in the Breast Cancer Association Consortium. International journal of epidemiology, 49(1), 216-232.

Keleş, A., Keleş, A., Yavuz, U. (2011). Expert system based on neuro-fuzzy rules for diagnosis breast cancer, Expert Systems with Applications, vol. 38, no. 5, pp. 5719-5726. Doi: 10.1016/j.eswa.2010.10.061

Keleş, A., Keleş, A. (2013). Extracting fuzzy rules for the diagnosis of breast cancer, Turkish Journal of Electrical Engineering & Computer Sciences, vol. 21, no. 5, pp. 1495-1503. doi:10.3906/elk-1012-938

Keleş, M. K., (2019). Breast cancer prediction and detection using data mining classification algorithms: a comparative study. *Tehničkivjesnik*, *26*(1), 149-155.

Khourdifi, Y., &Bahaj, M., (2018, December). Applying best machine learning algorithms for breast cancer prediction and classification. In *2018 International conference on electronics, control, optimization and computer science (ICECOCS)* (pp. 1-5). IEEE.

Kumar, C. A., Sooraj, M. P., & Ramakrishnan, S., (2017). A comparative performance evaluation of supervised feature selection algorithms on microarray datasets. *Procedia computer science*, *115*, 209-217.

Lévy, D., & Jain, A., (2016). Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*.

Lopez, M. G., Posada, N., Moura, D. C., Pollán, R. R., Valiente, J. M. F., Ortega, C. S., ... & Araújo, B. M. F. (2012). BCDR: a breast cancer digital repository. In 15th International conference on experimental mechanics (Vol. 1215).

Magna, A. A. R., Allende-Cid, H., Taramasco, C., Becerra, C., & Figueroa, R. L., (2020). Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis. *Ieee Access*, *8*, 106198-106213.

Maliha, S. K., Ema, R. R., Ghosh, S. K., Ahmed, H., Mollick, M. R. J., & Islam, T., (2019). Cancer disease prediction using Naive Bayes, K-nearest neighbor and J48 algorithm. In *2019 10th International Conference on*

Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

Maysanjaya, I. M. D., Pradnyana, I. M. A., &Putrama, I. M., (2018, June). Classification of breast cancer using Wrapper and Naïve Bayes algorithms. In Journal of Physics: Conference Series (Vol. 1040, No. 1, p. 012017). IOP Publishing.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic reviews, 4(1), 1-9.

Motlagh, M. H., Jannesari, M., Aboulkheyr, H., Khosravi, P., Elemento, O., Totonchi, M., &Hajirasouliha, I. (2018). Breast cancer histopathological image classification: A deep learning approach. *BioRxiv*, 242818.

Nakamura, Y., Hanaoka, S., Nomura, Y., Nakao, T., Miki, S., Watadani, T., ... & Abe, O., (2021). Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. *BMC Medical Informatics and Decision Making*, *21*(1), 1-19.

Negnevitsky, M. (2004). Design of a hybrid neuro-fuzzy decision-support system with a heterogeneous structure", 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542), vol. 2, pp. 1049-1052. Doi: 10.1109/FUZZY.2004.1375554

Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L. (2017). A knowledge-based system for breast cancer classification using fuzzy logic method, Telematics and Informatics, vol. 34, no. 4, pp. 133-144. Doi: 10.1016/j.tele.2017.01.007

Nguyen, E., Theodorakopoulos, D., Pathak, S., Geerdink, J., Vijlbrief, O., Van Keulen, M., & Seifert, C., (2020, October). A hybrid text classification and language generation model for automated summarization of dutch breast

cancer radiology reports. In 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) (pp. 72-81). IEEE.

Niknejad, M.T., (January 28, 2022). Breast imaging-reporting and data system (BI-RADS). *Radiopedia*.https://radiopaedia.org/articles/breast-imaging-reporting-and-data-system-bi-rads?lang=us

Nolan, M., (2022). What is a breast biopsy?,Verywellhealth. https://www.verywellhealth.com/open-surgical-breast-biopsy-429949.

Özçift, A., (2020). Medical Sentiment Analysis Based on Soft Voting Ensemble Algorithm. *YönetimBilişimSistemleriDergisi*, *6*(1), 42-50.

Öztürk, T., Aksoy, A., (2018). Turkish stopwords. https://github.com/ahmetax/trstop

Parlak, B., &Uysal, A. K. (2020). On classification of abstracts obtained from medical journals. *Journal of Information Science*, *46*(5), 648-663.

Parvaiz, A., Khalid, M. A., Zafar, R., Ameer, H., Ali, M., &Fraz, M. M., (2022). Vision Transformers in Medical Computer Vision--A Contemplative Retrospection. *arXiv preprint arXiv:2203.15269*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... &Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Prummel, M. V., Muradali, D., Shumak, R., Majpruz, V., Brown, P., Jiang, H., ... & Chiarelli, A. M., (2016). Digital compared with screen-film mammography: measures of diagnostic accuracy among women screened in the Ontario breast screening program. *Radiology*, *278*(2), 365-373.

Rajendran, K., Jayabalan, M., &Thiruchelvam, V., (2020). Predicting breast cancer via supervised machine learning methods on class imbalanced data. International Journal of Advanced Computer Science and Applications, 11(8).

Ramadevi, G. N., Rani, K. U., & Lavanya, D., (2015). Importance of feature extraction for classification of breast cancer datasets—a

study. *International Journal of Scientific and Innovative Mathematical Research*, *3*(2), 763-368.

Rao, R. R., &Makkithaya, K., (2017). Learning from a Class Imbalanced Public Health Dataset: a Cost-based Comparison of Classifier Performance. *International Journal of Electrical & Computer Engineering (2088-8708)*, *7*(4).

Ruiz, P., (2018). Understanding and visualizing resnets, Towards data science, https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8.

Sahran, S., Qasem, A., Omar, K., Albashih, D., Adam, A., Abdullah, S. N. H. S., ... & Abd Shukor, N. (2018). Machine learning methods for breast cancer diagnostic. *Breast Cancer and Surgery*.

Saib, W., Sengeh, D., Dlamini, G., & Singh, E., (2020). Hierarchical deep learning ensemble to automate the classification of breast cancer pathology reports by icd-o topography. *arXiv preprint arXiv:2008.12571*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T., (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sahria, I Mandang (2019). Diagnosis study of carcinoma mammae (breast cancer) disease using fuzzy logic method", In Journal of Physics: Conference Series, vol. 1277, no. 1, pp. 012039. doi:10.1088/1742-6596/1277/1/012039

Selvathi, D., &AarthyPoornila, A., (2018). Deep learning techniques for breast cancer detection using medical image analysis. In Biologically rationalized computing techniques for image processing applications (pp. 159-186). Springer, Cham.

Senthil Kumar, A.V. (Ed.) (2015). Fuzzy expert systems for disease diagnosis, pp. 200-246, IGI Global. Doi: 10.4018/978-1-4666-7240-6.

Shen, L. (2017). End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *arXiv preprint arXiv:1711.05775*.

Shin, B., Chokshi, F. H., Lee, T., & Choi, J. D., (2017, May). Classification of radiology reports using neural attention models. In *2017 international joint conference on neural networks (IJCNN)* (pp. 4363-4370). IEEE.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singer, G., &Marudi, M., (2020). Ordinal decision-tree-based ensemble approaches: The case of controlling the daily local growth rate of the COVID-19 epidemic. Entropy, 22(8), 871.

Singh, H., & Lone, Y. A. (2020). Deep neuro-fuzzy systems with python. Berkeley: Apress.

Sivakami, K., & Saraswathi, N. (2015). Mining big data: breast cancer prediction using DT-SVM hybrid model. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, *1*(5), 418-429.

Sizilio, G.R., Leite, C.R., Guerreiro, A.M., Neto, A.D.D. (2012). Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis, Biomedical engineering online, vol. 11, no. 1, pp. 1-21. Doi:10.1186/1475-925X-11-83.

Soui, M., Mansouri, N., Alhamad, R., Kessentini, M., &Ghedira, K., (2021). NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient's symptoms. Nonlinear dynamics, 106(2), 1453-1475.

Sree, S. V., Ng, E. Y. K., Acharya, R. U., & Faust, O. (2011). Breast imaging: a survey. *World journal of clinical oncology*, *2*(4), 171.

Suárez-Paniagua, V., Dong, H., & Casey, A., (2021). A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports. In *CLEF (Working Notes)* (pp. 846-856).

Suckling J, P. (1994). The mammographic image analysis society digital mammogram database. *Digital Mammo*, 375-386.

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., ... & Zhu, H. P., (2017). Risk factors and preventions of breast cancer. *International journal of biological sciences*, *13*(11), 1387.

Tokgoz, M., Turhan, F., Bolucu, N., & Can, B., (2021, February). Tuning Language Representation Models for Classification of Turkish News. In 2021 International Symposium on Electrical, Electronics and Information Engineering (pp. 402-407).

Turkish Bert Model. (2020). *Hugging face.*https://huggingface.co/dbmdz/bert-base-turkish-cased

Uzunhisarcıklı, E., Göreke, V., &Vekil, S. A. R. I., (2020). Comparison of type-2 fuzzy inference method and deep neural networks for mass detection from breast ultrasonography images. Cumhuriyet Science Journal, 41(4), 968-975.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I., (2017). Attention is all you need. Advances in neural information processing systems, 30.

Venkatesan, E., & Velmurugan, T., (2015). Performance analysis of decision tree algorithms for breast cancer classification. Indian Journal of Science and Technology, 8(29), 1-8.

Verma, A., (2021). Python Guide to HuggingFaceDistilBERT – Smaller, Faster & Cheaper Distilled BERT

Wang, M., & Hu, F., (2021). The Application of NLTK Library for Python Natural Language Processing in Corpus Research. *Theory and Practice in Language Studies*, *11*(9), 1041-1049.

Wei, J., & Zou, K., (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M., (2005, June). Practical machine learning tools and techniques. In Data Mining (Vol. 2, No. 4).

World Health Organization, (2020). Breast [Online] Available: https://www.gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf. Accessed [13 September 2021].

Yadav, S. S., & Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, *6*(1), 1-18.

Yan, A., McAuley, J., Lu, X., Du, J., Chang, E. Y., Gentili, A., & Hsu, C. N., (2022). RadBERT: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, *4*(4), e210258.

Yogarajan, V., Montiel, J., Smith, T., &Pfahringer, B., (2021, June). Transformers for multi-label classification of medical text: an empirical comparison. In *International Conference on Artificial Intelligence in Medicine* (pp. 114-123). Springer, Cham.).

Youn, H. J., & Han, W., (2020). A review of the epidemiology of breast cancer in Asia: focus on risk factors. *Asian Pacific journal of cancer prevention: APJCP*, *21*(4), 867.

Zadeh, L. A. (1988). Fuzzy logic. Computer, 21(4), 83-93.

Zand, H. K. K. (2015). A comparative survey on data mining techniques for breast cancer diagnosis and prediction. *Indian Journal of Fundamental and Applied Life Sciences*, *5*(s1), 4330-9.

Zorluoglu, G., &Agaoglu, M. (2017). Diagnosis of breast cancer using ensemble of data mining classification methods. *International Journal of Oncology and Cancer Therapy*, *2*.

BIOGRAPHY

Pınar USKANER HEPSAĞ. She received his BSc. from Mathematics Department of Cukurova University in 2008, MSc from Computer Science Department of Case Western Reserve University in 2012.Now, she has been working as LectureratAdanaAlparslanTurkesScience and Technology University since 2014. Her research interests include Image processing, Machine Learning, Deep Learning, and Transformer models.

APPENDICES

A. The actual label and predicted label of each patient for image, survey, and text.

Table A.1. The prediction results of each model.

Test instances	**Actual label**	**Survey_predicted label with atrributeselectedclassifier**	**Image_predicted label with VGG16**	**Text_predicted label with Transformer**
1	malign	malign	malign	benign
2	malign	malign	malign	malign
3	malign	benign	malign	malign
4	malign	malign	malign	malign
5	malign	malign	malign	benign
6	benign	malign	benign	benign
7	benign	benign	benign	benign
8	benign	benign	benign	benign
9	benign	benign	benign	benign
10	benign	benign	benign	benign
11	benign	benign	benign	benign
12	benign	benign	benign	benign
13	benign	benign	benign	benign

B. Different preprocessingcombinations applied to Survey Dataset.

Augmentation+Normalization+Classification

As a first step, we applied three different augmentation methods to the training data and then normalized the data before the classification process, as shown in Table B.1, Table B.2, and Table B.3. Table B.1, Table B.2, and Table B.3show the classification results with different algorithms for different training data obtained with different augmentation methods.

Table B.1 shows the classification results obtained by tripling the training data and then using the algorithm SMOTE to balance the augmented data. We then applied the MinMax method to normalize the data.

Table B.1 Train Tripled+ SMOTE + MinMax + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	**0.81**	**0.3**	**0.83**	**0.81**	**0.82**
attributeselected	0.78	0.36	0.81	0.78	0.79
bayesnet	0.74	0.22	0.82	0.74	0.78
Cnn	0.61	0.42	0.66	0.61	0.63
Bagging	0.80	0.33	0.82	0.80	0.81
j48	0.63	0.58	0.58	0.63	0.60
Jrip	0.68	0.40	0.68	0.68	0.68
Libsvm	0.7	0.46	0.56	0.7	0.62
naivebayes	0.71	0.26	0.78	0.71	0.74
Oner	0.55	0.54	0.57	0.55	0.56
randomforest	0.73	0.43	0.61	0.73	0.66
randomtree	0.6	0.54	0.57	0.6	0.58

Table B.2Train+ SMOTE + MinMax + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	**0.81**	**0.25**	**0.84**	**0.81**	**0.82**
Attributeselected	0.76	0.46	0.79	0.76	0.77
bayesnet	0.65	0.29	0.73	0.65	0.69
Cnn	0.71	0.34	0.77	0.71	0.74
Bagging	0.78	0.28	0.87	0.78	0.82
j48	0.71	0.39	0.72	0.71	0.71
Jrip	0.68	0.46	0.68	0.68	0.68
Libsvm	0.68	0.39	0.76	0.68	0.72
naivebayes	0.69	0.24	0.79	0.69	0.74
Oner	0.73	0.41	0.75	0.73	0.74
randomforest	0.78	0.34	0.80	0.78	0.79
randomtree	0.68	0.46	0.68	0.68	0.68

Table B.3Tripled+ MinMax + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	0.81	0.33	0.85	0.81	0.83
Attributeselected	**0.83**	**0.33**	**0.85**	**0.83**	**0.84**
bayesnet	0.76	0.2	0.81	0.76	0.78
Cnn	0.66	0.37	0.71	0.66	0.68
Bagging	0.79	0.36	0.82	0.79	0.80
j48	0.65	0.6	0.55	0.65	0.60
Jrip	0.65	0.52	0.62	0.65	0.63
Libsvm	0.75	0.45	0.76	0.75	0.75
naivebayes	0.76	0.2	0.81	0.76	0.78
Oner	0.55	0.38	0.63	0.55	0.59
randomforest	0.79	0.34	0.69	0.79	0.74
randomtree	0.82	0.26	0.85	0.82	0.83

Table B.2 and Table B.3 show the classification results obtained with the training data using the algorithm SMOTE without augmentation and by tripling the training data and balancing with self-copy without SMOTE, respectively. The MinMax method was then applied to normalize the data.

Augmentation+Normalization+FeatureSelection+Classification

After the augmentation and normalization processes were completed, attribute selection was performed using CfsSubsetEval. Once the attributes were determined, the classification process was performed. Table B.4, Table B.5, and Table B.6 show the classification results of the different algorithms for CFSSubsetEval feature selector.

Table B.4Train Tripled+ SMOTE+MinMax +CFSSubsetEval+ Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	0.55	0.47	0.57	0.55	0.55
Attributeselected	0.73	0.48	0.81	0.73	0.67
bayesnet	0.46	0.63	0.46	0.46	0.46
Cnn	0.46	0.63	0.46	0.46	0.46
Bagging	**0.73**	**0.48**	**0.81**	**0.73**	**0.77**
j48	0.73	0.48	0.81	0.73	0.77
Jrip	0.55	0.38	0.63	0.55	0.59
Libsvm	0.46	0.63	0.46	0.46	0.46
naivebayes	0.46	0.63	0.46	0.46	0.46
Oner	0.46	0.53	0.50	0.46	0.48
randomforest	**0.73**	**0.48**	**0.81**	**0.73**	**0.77**
randomtree	0.55	0.69	0.38	0.55	0.43

Table B.5Train+SMOTE+MinMax +CFSSubsetEval+ Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	0.74	0.42	0.76	0.74	0.75
Attributeselected	0.67	0.48	0.65	0.67	0.66
bayesnet	0.74	0.39	0.76	0.74	0.75
Cnn	**0.74**	**0.36**	**0.77**	**0.74**	**0.75**
Bagging	0.65	0.49	0.64	0.65	0.64
j48	0.67	0.48	0.65	0.67	0.66
Jrip	0.68	0.51	0.66	0.68	0.67
Libsvm	0.65	0.49	0.64	0.65	0.64
naivebayes	0.69	0.38	0.73	0.69	0.71
Oner	0.73	0.41	0.75	0.73	0.74
randomforest	0.7	0.49	0.69	0.7	0.69
randomtree	0.66	0.51	0.69	0.66	0.67

Table B.6. Tripled+MinMax +CFSSubsetEval+ Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	0.60	0.46	0.58	0.60	0.59
Attributeselected	0.65	0.48	0.64	0.65	0.64
bayesnet	0.73	0.40	0.77	0.73	0.75
Cnn	0.63	0.46	0.6	0.63	0.61
Bagging	0.63	0.53	0.57	0.63	0.60
j48	0.65	0.48	0.64	0.65	0.64
Jrip	0.65	0.51	0.59	0.65	0.62
Libsvm	0.63	0.53	0.57	0.63	0.60
naivebayes	**0.80**	**0.26**	**0.81**	**0.80**	**0.80**
Oner	0.42	0.57	0.47	0.42	0.44
randomforest	0.68	0.50	0.69	0.68	0.68
randomtree	0.63	0.51	0.61	0.63	0.62

Augmentation+Standardization+WEKA classifiers:

In this part, we first augmented the data and then standardized the data using the standard scaler method. Table B.7, Table B.8, and Table B.9 show the classification results with different algorithms for augmented data with standardization.

Table B.7Train Tripled+ SMOTE + StandardScaler + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	0.79	0.31	0.82	0.79	0.80
Attributeselected	0.76	0.38	0.81	0.76	0.78
bayesnet	0.72	0.23	0.79	0.72	0.76
Cnn	0.58	0.47	0.62	0.58	0.60
Bagging	**0.83**	**0.28**	**0.86**	**0.83**	**0.84**
j48	0.62	0.57	0.59	0.62	0.60
Jrip	0.6	0.54	0.59	0.6	0.59
Libsvm	0.71	0.45	0.76	0.71	0.73
naivebayes	0.74	0.21	0.80	0.74	0.77
Oner	0.55	0.54	0.57	0.55	0.56
randomforest	0.74	0.41	0.74	0.74	0.74
randomtree	0.7	0.43	0.7	0.7	0.70

Table B.8Train + SMOTE + StandardScaler + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
adaboost	0.75	0.44	0.73	0.75	0.74
Attributeselected	0.72	0.47	0.75	0.72	0.73
bayesnet	0.61	0.32	0.72	0.61	0.66
Cnn	0.62	0.42	0.72	0.62	0.67
Bagging	**0.82**	**0.33**	**0.83**	**0.82**	**0.82**
j48	0.60	0.60	0.59	0.60	0.59
jrip	0.65	0.44	0.62	0.65	0.63
libsvm	0.61	0.46	0.68	0.61	0.64
naivebayes	0.64	0.28	0.76	0.64	0.69
Oner	0.72	0.45	0.61	0.75	0.67
randomforest	0.66	0.52	0.68	0.66	0.67
randomtree	0.60	0.45	0.66	0.60	0.63

Table B.9 Tripled + StandardScaler + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	0.79	0.33	0.81	0.79	0.80
Attributeselected	**0.81**	**0.38**	**0.83**	**0.81**	**0.82**
Bayesnet	0.76	0.2	0.81	0.76	0.78
Cnn	0.66	0.37	0.71	0.66	0.68
Bagging	0.79	0.36	0.82	0.79	0.80
j48	0.68	0.56	0.59	0.68	0.63
Jrip	0.60	0.50	0.61	0.60	0.60
Libsvm	0.72	0.44	0.72	0.72	0.72
naivebayes	0.79	0.16	0.84	0.79	0.81
Oner	0.57	0.37	0.65	0.57	0.61
randomforest	0.71	0.46	0.68	0.71	0.69
randomtree	0.65	0.48	0.66	0.65	0.65

Normalization+Augmentation +WEKA classifiers:

We first apply the normalization method and then expand the normalized data to check the classification results for different algorithms. Table B.10, Table B.11, and Table B.12 show the classification results with different algorithms for normalized and expanded data.

Table B.10MinMax +Train Tripled+SMOTE + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	0.72	0.35	0.78	0.72	0.75
Attributeselected	0.76	0.33	0.79	0.76	0.77
Bayesnet	0.63	0.50	0.56	0.63	0.59
Cnn	0.76	0.31	0.80	0.76	0.78
Bagging	**0.79**	**0.27**	**0.82**	**0.79**	**0.81**
j48	0.76	0.34	0.78	0.76	0.77
Jrip	0.71	0.39	0.73	0.71	0.72
Libsvm	0.58	0.54	0.60	0.58	0.59
naivebayes	0.73	0.38	0.72	0.73	0.72
Oner	0.47	0.64	0.50	0.47	0.48
randomforest	0.66	0.46	0.70	0.66	0.68
randomtree	0.68	0.47	0.67	0.68	0.67

Table B.11MinMax+ Train+SMOTE + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	0.70	0.36	0.71	0.70	0.70
attributeselected	0.72	0.33	0.76	0.72	0.74
Bayesnet	0.57	0.37	0.67	0.57	0.62
Cnn	**0.75**	**0.38**	**0.78**	**0.75**	**0.77**
Bagging	0.64	0.46	0.69	0.64	0.66
j48	0.63	0.46	0.65	0.63	0.64
Jrip	0.68	0.43	0.75	0.68	0.71
Libsvm	0.59	0.57	0.64	0.59	0.61
naivebayes	0.66	0.44	0.66	0.66	0.66
Oner	0.56	0.59	0.58	0.56	0.57
randomforest	0.64	0.46	0.69	0.64	0.66
randomtree	0.68	0.49	0.68	0.68	0.68

Table B.12MinMax +Tripled + Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	**0.76**	**0.32**	**0.72**	**0.78**	**0.75**
attributeselected	0.71	0.44	0.71	0.71	0.71
Bayesnet	0.65	0.70	0.56	0.65	0.60
Cnn	0.66	0.31	0.74	0.66	0.70
Bagging	0.71	0.43	0.71	0.71	0.71
j48	0.70	0.41	0.72	0.70	0.71
Jrip	0.69	0.43	0.54	0.69	0.61
Libsvm	0.58	0.51	0.61	0.58	0.59
naivebayes	0.66	0.62	0.63	0.66	0.64
Oner	0.55	0.61	0.55	0.55	0.55
randomforest	0.67	0.52	0.66	0.67	0.66
randomtree	0.74	0.38	0.75	0.74	0.74

Augmentation+ Normalization +PCA+WEKA classifiers:

In this experiment, after augmentation and normalization, we applied dimensionality reduction with PCA instead of feature selection. The classification results of the different algorithms obtained from PCA with augmented and normalized data are shown in in Table B.13, Table B.14, and Table B.15.

Table B.13.Train Tripled+ SMOTE + MinMax + PCA+Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	0.72	0.35	0.78	0.72	0.75
attributeselected	0.76	0.33	0.79	0.76	0.77
Bayesnet	0.63	0.50	0.63	0.63	0.68
Cnn	0.76	0.31	0.8	0.76	0.78
Bagging	**0.79**	**0.27**	**0.82**	**0.79**	**0.80**
j48	0.76	0.34	0.78	0.76	0.77
Jrip	0.71	0.39	0.73	0.71	0.72
Libsvm	0.58	0.54	0.6	0.58	0.59
naivebayes	0.73	0.38	0.72	0.73	0.73
Oner	0.47	0.64	0.5	0.47	0.48
randomforest	0.66	0.46	0.70	0.66	0.68
randomtree	0.68	0.47	0.67	0.68	0.67

Table B.14Train+SMOTE + MinMax+ PCA+Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	0.7	0.36	0.71	0.7	0.7
attributeselected	0.72	0.43	0.72	0.72	0.72
Bayesnet	0.57	0.41	0.64	0.57	0.60
Cnn	**0.75**	**0.38**	**0.78**	**0.75**	**0.77**
Bagging	0.66	0.48	0.70	0.66	0.68
j48	0.63	0.46	0.65	0.63	0.64
Jrip	0.68	0.43	0.75	0.68	0.71
Libsvm	0.59	0.57	0.64	0.59	0.61
naivebayes	0.66	0.44	0.66	0.66	0.66
Oner	0.56	0.59	0.57	0.56	0.56
randomforest	0.66	0.48	0.70	0.66	0.68
randomtree	0.68	0.46	0.68	0.68	0.68

Table B.15Tripled + MinMax +PCA+ Classification using WEKA.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	**0.76**	**0.32**	**0.78**	**0.76**	**0.77**
attributeselected	0.71	0.45	0.71	0.71	0.71
Bayesnet	0.65	0.70	0.56	0.65	0.60
Cnn	0.66	0.36	0.74	0.66	0.70
Bagging	0.71	0.43	0.71	0.71	0.71
j48	0.70	0.41	0.72	0.70	0.71
Jrip	0.69	0.43	0.54	0.69	0.61
Libsvm	0.58	0.51	0.61	0.58	0.59
naivebayes	0.66	0.55	0.63	0.66	0.64
Oner	0.55	0.61	0.55	0.55	0.55
randomforest	0.69	0.52	0.68	0.69	0.68
randomtree	0.74	0.38	0.75	0.74	0.74

WithoutAugmentation+WithoutNormalization+WithoutFeatureSelector+WEKA classifiers:

In this part, we use data without applying augmentation, normalization, and feature selection methods. We test pure data using different classification algorithms. Table B.16 shows the weighted average results of 5-folds for all classification algorithms.

Table B.16. Using pure data without augmentation+withoutNormalization+without Feature Selector.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure
Adaboost	0.76	0.40	0.77	0.76	0.76
Attributeselected	**0.84**	**0.2**	**0.88**	**0.84**	**0.86**
Bayesnet	0.75	0.47	0.74	0.75	0.74
Cnn	0.68	0.36	0.73	0.68	0.70
Bagging	0.73	0.49	0.75	0.73	0.74
j48	0.74	0.38	0.78	0.74	0.76
Jrip	0.78	0.35	0.82	0.78	0.80
Libsvm	0.71	0.41	0.73	0.71	0.72
Naivebayes	0.79	0.3	0.84	0.79	0.81
Oner	0.68	0.51	0.70	0.68	0.69
randomforest	0.78	0.32	0.79	0.78	0.78
Randomtree	0.67	0.43	0.66	0.67	0.66