

An AI-based Approach to Academic Quality Quantification

A thesis submitted to the
Institute for Data Science and Artificial Intelligence

by

ZEYNEP KARAMAN

in partial fulfillment for the
degree of Master of Science

in

Data Science and Artificial Intelligence



Declaration of Authorship

I, ZEYNEP KARAMAN, declare that this thesis titled, "An AI-based Approach to Academic Quality Quantification" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 08.08.2025

An AI-based Approach to Academic Quality Quantification

ZEYNEP KARAMAN

Abstract

In the context of academic recruitment at universities and research institutions, establishing consistent and effective evaluation criteria remains a complex challenge. Identifying robust metrics that align with globally recognized standards of academic quality is essential to ensure merit-based evaluation of researchers and maintain institutional credibility. In this thesis, we address this challenge through an AI-based solution aimed at developing a data-driven approach to quantify academic quality. As a benchmark of academic excellence, we use Nobel laureates' profiles in Physics, Chemistry, Physiology or Medicine, and Economics as a reference cohort. Comparison group includes researchers from the same fields affiliated with universities with an average ranking based on the Times Higher Education World University Rankings. By defining bibliometric features of the academic profiles of both Nobel laureates and the comparison group, we develop machine learning models to quantify academic quality and identify key features that define academic excellence. The ultimate goal is to support the decision-making process of universities and research institutions in academic recruitment, creating fair and objective evaluation criteria.

Keywords: Data Science, Machine Learning, Scientometrics, Academic Quality Quantification, Recommendation Systems

Akademik Kalite Ölçümü İçin Yapay Zeka Tabanlı Bir Yaklaşım

ZEYNEP KARAMAN

ÖZ

Üniversiteler ve araştırma kurumlarında akademik değerlendirme ve işe alım bağlamında tutarlı ve etkili değerlendirme kriterleri oluşturmak karmaşık bir konudur. Araştırmacıların liyakate dayalı şekilde değerlendirilmesi ve kurumsal güvenilirliği sürdürmek için, akademik kalite standartlarının belirlenmesi büyük önem taşımaktadır.

Bu tez çalışmasında, akademik kaliteyi nicel olarak değerlendirmeye yönelik yapay zeka tabanlı bir çözüm öneriyoruz. Akademik kalitenin bir ölçütü olarak, Fizik, Kimya, Fizyoloji veya Tıp, ve Ekonomi alanlarında Nobel ödülü kazanmış araştırmacıların profilleri referans grubunu oluşturmaktadır. Karşılaştırma grubu, aynı alanlarda çalışan ve Times Higher Education Dünya Üniversite Sıralamaları'na göre ortalama sıralamaya sahip üniversitelerde görev yapan araştırmacılardan oluşmaktadır.

Nobel ödüllü araştırmacılar ile karşılaştırma grubundaki araştırmacıların akademik profillerine ilişkin bibliyometrik metrikler tanımlanmıştır. Bu metrikler kullanılarak, akademik kaliteyi nicel olarak ölçmek ve akademik kaliteyi tanımlayan temel göstergeleri belirlemek amacıyla makine öğrenmesi modelleri geliştirilmiştir. Nihai amaç, üniversiteler ve araştırma kurumlarının akademik işe alım süreçlerinde karar verme mekanizmalarını destekleyerek, objektif değerlendirme kriterleri oluşturulmasına katkı sağlamaktır.

Anahtar Sözcükler: Veri Bilimi, Makine Öğrenmesi, Bilimmetri, Akademik Kalite Ölçümü, Öneri Sistemleri

Acknowledgments

I would like to sincerely thank my thesis advisor, **Dr. Ercan Atam**, for his extensive support, guidance, and motivation. His insightful feedback and considerate mentorship were invaluable at every stage of this thesis. I am thankful for the opportunity to work on this topic under his supervision.

I would also like to thank my thesis co-advisor, **Dr. Şaziye Betül Özateş**, for her thoughtful guidance, constructive feedback, and willingness to support me. Her insights and perspective significantly enriched the quality of this research.

Lastly, I would like to thank the thesis committee members, **Dr. Hüseyin Oktay Altun**, **Dr. Erhan Yumuk**, and **Dr. Tarkan Aydın** for their time and participation in evaluation.

Contents

Declaration of Authorship	ii
Abstract	iii
Öz	iv
Acknowledgments	v
List of Figures	viii
List of Tables	ix
Abbreviations	x
Symbols	xi
1 Introduction	1
1.1 Introduction	1
1.2 Historical background	2
1.3 Original value	5
2 Methodology	7
2.1 Data description	7
2.2 Data collection	7
2.3 Data preprocessing	8
2.4 Features for academic quality quantification	9
2.5 Model selection	13
3 Models for Academic Quality Quantification	15
3.1 Siamese neural network model	15
3.1.1 Overview of Siamese neural networks	15
3.1.2 Siamese neural network for academic quality quantification	17
3.1.2.1 Siamese neural network architecture	18
3.1.2.2 Loss function: Modified contrastive loss	21
3.1.2.3 Hyperparameters of Siamese neural network	22
3.1.2.4 Final model training and testing	23
3.2 Ensemble learning model	24
3.2.1 Overview of XGBoost	24
3.2.2 XGBoost for academic quality quantification	25

3.2.2.1	XGBoost architecture	25
3.2.2.2	Loss function	29
3.2.2.3	Hyperparameters of XGBoost	29
3.2.2.4	Final model training and testing	31
3.3	Feature importance determination	31
3.3.1	Feature importance for Siamese neural network	31
3.3.2	Feature importance for ensemble learning	33
4	Results	35
4.1	Results of all-field data	35
4.2	Results of field-based data	37
4.2.1	Physics	37
4.2.2	Chemistry	39
4.2.3	Physiology or Medicine	41
4.2.4	Economics	43
4.3	Comparison of models	45
4.4	Feature Importances	46
4.5	Discussion of results	47
4.5.1	Summary of findings	47
4.5.2	Interpretation	49
4.5.3	Implications	52
4.5.4	Limitations	52
5	Conclusion and Future Directions	55
A	University List	57
A.1	List of universities for the average class researchers	57
B	Pseudocode	61
B.1	Siamese neural network model	61
B.2	XGBoost model	67
	Bibliography	70

List of Figures

3.1	Siamese NN model.	16
3.2	Siamese NN for AQI quantification.	18
3.3	XGBoost model for AQI quantification (adapted from [1]).	28
4.1	AQI distribution for Siamese NN model (training) - All Fields.	35
4.2	AQI distribution for Siamese NN model (test) - All Fields.	36
4.3	Training loss and MQ score for Siamese NN model - All Fields.	36
4.4	AQI distribution for XGBoost model (training) - All Fields.	36
4.5	AQI distribution for XGBoost model (test) - All Fields.	37
4.6	AQI distribution for Siamese NN model (training) - Physics.	37
4.7	AQI distribution for Siamese NN model (test) - Physics.	38
4.8	Training loss and MQ score for Siamese NN model - Physics.	38
4.9	AQI distribution for XGBoost model (training) - Physics.	38
4.10	AQI distribution for XGBoost model (test) - Physics.	39
4.11	AQI distribution for Siamese NN model (training) - Chemistry.	39
4.12	AQI distribution for Siamese NN model (test) - Chemistry.	40
4.13	Training loss and MQ score for Siamese NN model - Chemistry.	40
4.14	AQI distribution for XGBoost model (training) - Chemistry.	40
4.15	AQI distribution for XGBoost model (test) - Chemistry.	41
4.16	AQI distribution for Siamese NN model (training) - Medicine.	41
4.17	AQI distribution for Siamese NN model (test) - Medicine.	42
4.18	Training loss and MQ score for Siamese NN model - Medicine.	42
4.19	AQI distribution for XGBoost model (training) - Medicine.	42
4.20	AQI distribution for XGBoost model (test) - Medicine.	43
4.21	AQI distribution for Siamese NN model (training) - Economics.	43
4.22	AQI distribution for Siamese NN model (test) - Economics.	44
4.23	Training loss and MQ score for Siamese NN model - Economics.	44
4.24	AQI distribution for XGBoost model (training) - Economics.	44
4.25	AQI distribution for XGBoost model (test) - Economics.	45
4.26	Feature importance scores for XGBoost and Siamese NN models - All Fields.	47
4.27	Feature importance scores for XGBoost model - Physics and Chemistry.	47
4.28	Feature importance scores for XGBoost model - Medicine and Economics.	47

List of Tables

2.1	Data description.	8
2.2	Features for academic quality quantification.	9
3.1	Advantages and challenges of Siamese NN.	17
3.2	Training parameters for the Siamese NN model.	23
3.3	Advantages and challenges of XGBoost.	25
3.4	Training parameters for the XGBoost model.	30
3.5	Feature importance determination vs. Feature selection.	31
4.1	Comparison of training and test performance of models based on MQ score - All Fields.	45
4.2	Comparison of training and test performance of models based on MQ score - Physics.	45
4.3	Comparison of training and test performance of models based on MQ score - Chemistry.	45
4.4	Comparison of training and test performance of models based on MQ score - Physiology or Medicine.	46
4.5	Comparison of training and test performance of models based on MQ score - Economics.	46
4.6	Feature importances for the best-performing model.	46
4.7	Top 5 feature importance rankings.	46
4.8	Summary of test results for Siamese NN and XGBoost models.	48
5.1	Summary comparison between the Siamese NN and XGBoost.	56
A.1	List of universities for the average class researchers.	57

Abbreviations

AI	A rtificial I ntelligence
AQI	A cademic Q uality I ndex
CV	C ross V alidation
L1	L1 regularization (Lasso)
L2	L2 regularization (Ridge)
MQ	M odel Q uality
NN	N eural N etwork
SDG	S tochastic G radient D escent
SVM	S upport V ector M achine
XGBoost	eX treme G radient B oosting
WoS	W eb o f S cience

Symbols

Symbol Name

a_i	Author position in the co-author list of publication
b	Bias
Cit	Number of citations
G	Gradient sums (first derivatives of loss) for child nodes of the ensemble learning model
H	Hessian sums (second derivatives of loss) for child nodes of the ensemble learning model
$L_{\text{contrastive}}$	Contrastive loss term
L_{mcl}	Modified contrastive loss function
L_{order}	Order loss term
m_c	Contrastive loss margin
m_n	Mean AQI values for the negative class
m_o	Order loss margin
m_p	Mean AQI values for the positive class
P	Probability
Pub	Publications
Pub_{count}	Number of publications
r_i	Residual (difference between the true label and prediction of instance x_i)
s_n	Variance of AQI values within the negative class
s_p	Variance of AQI values within the positive class
W	Weight
x_i	Input vector of the i th researcher profile
y_i	True label of an instance x_i
$\hat{y}_i$	Predicted label of an instance x_i
e	Epsilon
$\log$	Logarithm

∂	Derivative
η	Learning rate
γ	Regularization parameter on the number of splits for ensemble learning model
λ	L2 regularization parameter
σ	Sigmoid
$\phi(\cdot)$	Activation function for the neural network model
$\Omega(\cdot)$	Regularization term for the ensemble learning model
$\sum$	Summation



Chapter 1

Introduction

1.1 Introduction

Universities and research institutions are the fundamental pillars of higher education, research, innovation, and human capital development. To successfully achieve their role, maintain progress and remain competitive, the ability to attract and retain qualified academic personnel is crucial. While selecting the best candidates for research or professorial positions is challenging, the task has become more difficult due to the increasing number of qualified applicants with diverse academic backgrounds.

One of the challenges in academic hiring is the absence of objective, data-driven and verifiable assessment methods. Expert panels and peer review are typically used to assess candidates, but this process is susceptible to subjectivity, inconsistency, and biases. This is especially true when committee members have different opinions about evaluation criteria or when decisions are heavily influenced by unofficial factors such as institutional inbreeding or personal networks.

The purpose of this thesis is to develop an indicator of academic quality for researchers to address this problem. Specifically, our objective is to develop an *Academic Quality Index (AQI)* to predict academic quality by developing two machine learning frameworks: Siamese neural network (NN) [2] and ensemble learning model.

The input for both models is composed of features that characterize researcher's output and impact, such as h-index, a-index, g-index, number of publications, and total citations.

The features are derived from bibliometric data for each researcher (e.g., number of publications, number of citations per publication, average author position per publication), all of which are stored in bibliographic databases. The output of both models is composed of the AQI and feature importance scores, which highlight the key factors that distinguish the overall academic quality of a researcher.

Both models were trained on data from researchers in four fields - Physics, Chemistry, Physiology or Medicine, and Economics. The benchmark group consists of Nobel laureates in these four fields, while the comparison group includes researchers affiliated with universities ranked between 601 and 800 in the Times Higher Education World University Rankings [3], which we consider to represent an average ranking.

The resulting tool can serve as a decision support mechanism for universities and research institutions, allowing a fair and consistent evaluation of academics for recruitment.

1.2 Historical background

Scientometrics is the field that analyzes how scientific knowledge is produced, shared and evaluated to understand the mechanisms of research [4]. Within this field, bibliometrics plays a central role by applying statistical methods to analyze publication and citation patterns. Bibliometric indicators are widely used to inform national research policies, guide funding decisions, support hiring and promotion processes, and assess the overall impact of research. Traditionally, peer review has served as the dominant method of evaluating scientific output and academic personnel, while bibliometric analysis provides a quantitative complement to this approach. When integrated with expert judgment, this approach - known as *informed peer review* - can improve quality of the evaluation, thus reducing bias and increasing transparency [5].

Bibliometrics can be categorized into two categories: descriptive and evaluative. Descriptive bibliometrics involves tracking metrics such as the number of publications, while evaluative bibliometrics focuses on assessing the impact of research, most commonly through citation analysis [6]. Among these, citation counts are generally regarded as a more reliable indicator of research influence [7]. However, simple counts of publications and citations have limitations. To address these, the scientific community has developed a wide range of specialized indices, including the h-index [8], g-index [9], a-index [10],

r-index [11], w-index [12], hg-index [13], m-index [14], e-index [15], and hi-index [16]. In fact, over 100 bibliometric indicators are used to evaluate research performance at the author level [17]. These metrics can be broadly grouped into two categories: research output and productivity measures (e.g., total publications, total citations, number of authors per paper, number of citations per year, number of citations per paper) and citation intensity-based indices (e.g., h-index, g-index, a-index, f-index, r-index, hg-index) [18]. Most of these indicators are derived from mathematical functions that incorporate factors such as publication count, citation frequency and the number of co-authors.

No single indicator can fully capture an individual's scientific impact, which makes a combination of metrics necessary for a valid evaluation [19]. Therefore, bibliometric research generally emphasizes the need for diverse metrics to evaluate different facets of a researcher's work [20]. Each indicator reflects a specific aspect of research performance, and the use of multiple indicators, ideally along with peer review, offers a more comprehensive assessment [21, 22].

Several studies have investigated the most relevant bibliometric indicators using traditional statistical methods such as correlation analysis. For example, a study of senior academics in Italy's occupational medicine field (based on citation analysis of products published during 2001-2010) highlighted three core metrics: number of publications, number of citations, and h-index [23]. Another study introduced the C-score, a composite metric that combines indicators such as citation count, h-index, hm-index (adjusted for co-authorship) and citations weighted by author position [24]. Clustering techniques have also been used to group author-level indicators and explore their relation to performance [25]. One study clustered 44 metrics that cover productivity, impact, and influence to categorize researchers by performance level [26]. Recent works in [27, 28] have also examined the relationship between traditional bibliometrics and altmetrics, sometimes integrating peer review as a quality measure.

Machine learning approaches have further advanced the evaluation of bibliometric indicators. A study applied a binomial regression model (logit) to data from 3,659 researchers across disciplines, using eight variables including variants of the h-index, publication and citation counts, and demographics, to predict promotion results [29]. Another study analyzed 363 biomedical Ph.D. graduates to predict tenure-track success using logistic regression and SVM, evaluating 10 indicators such as publication counts, authorship

roles, impact factors, and weighted outputs [30]. Logistic regression was also used to identify h-index variants that align with qualitative assessments by international civil engineering societies for award nominations [18]. A recent study applied Random Forest algorithms to evaluate 30 bibliometric features among mathematicians [31], while another used SVM to predict outcomes from Italy's National Scientific Habilitation, identifying key predictors for permanent faculty applications [32].

A substantial body of literature, including the studies referenced above, has analyzed bibliometric data to develop research assessment indicators to rank researchers, typically emphasizing metrics such as publication and citation counts. Previous studies also highlight a positive correlation between peer review outcomes and bibliometric indicators, reinforcing their value as complementary tools in academic evaluation. However, there is still no consensus in the academic community on the most significant features or metrics to assess academic quality. Accordingly, our review identified two primary gaps in the existing literature:

1. **Limited used of machine learning approaches:** Many of these studies rely on conventional ranking methods rather than leveraging modern machine learning techniques to evaluate the relative effectiveness of these indicators. There remains a need to assess and rank research indicators using contemporary machine learning approaches.
2. **Lack of comprehensive models, new features and data in existing machine learning studies:** Among the few studies that employ machine learning, most focus on specific fields, narrow data or feature sets, and a single model, without offering a generalizable approach. Two main limitations are: i) Limited disciplinary scope: Most studies are confined to a single academic field or lack clear reference groups for comparison. Although some research has attempted to evaluate the effectiveness of various indices, the absence of a standardized dataset limits the ability to make meaningful comparisons; and ii) Narrow methodological focus: Previous studies often consider only a small subset of bibliometric indicators and typically rely on a single model to assess research performance.

1.3 Original value

As a result, there is a lack of studies that simultaneously incorporate diverse bibliometric indicators, multiple machine learning models, field-spanning comparisons, and integrated datasets from multiple sources. This thesis aims to address these gaps by:

- Developing a new indicator called *Academic Quality Index* to assess the academic quality of a given researcher based on multiple features;
- Incorporating a wide array of bibliometric indices and introducing new metrics (*cpppp-index*, *aap-index* and *max_fa_citations*) as input features to capture multiple dimensions of researcher performance;
- Applying Siamese NN and ensemble learning models to measure AQI and feature importance scores to identify the most significant features for measuring academic quality;
- Covering four disciplines, namely Physics, Chemistry, Physiology and Medicine, and Economics;
- Using Nobel laureates as a benchmark group;
- Using professors affiliated with average ranking universities as a comparison group; and
- Combining data from multiple databases (e.g., Web of Science and Scopus) to create a robust dataset that is intended to be made publicly available.

By embedding bibliometric data within machine learning frameworks, the proposed AQI and feature importance scores offer predictive power for evaluation. We aim for our findings to complement peer review in the researcher evaluation process by offering objective and data-driven insights.

The following chapters are organized as follows. Chapter 2 provides details on the data description and collection process, and introduces the features used in the models to quantify academic quality. Chapter 3 presents two machine learning frameworks: Siamese NN and ensemble learning. The chapter then discusses the determination of feature importance scores in academic quality quantification. Chapter 4, presents the

results, including a comparison of the performance of models and a comprehensive discussion. Finally, Chapter 5 concludes the thesis and outlines the directions for future research.



Chapter 2

Methodology

2.1 Data description

In this thesis, we define two distinct academic classes to quantify academic quality. The benchmark class, referred to as the "positive class" (label=1), consists of Nobel laureates in four fields, namely Physics, Chemistry, Physiology or Medicine, and Economics. The second class, referred to as the "negative class" (label=0), consists of professors in the same four fields, affiliated with institutions ranked between 601 and 800 according to the 2025 Times Higher Education World University Rankings. We included only full professors from these universities for comparability purposes with Nobel laureates. These two classes serve as comparative baselines for training and evaluating machine learning models.

2.2 Data collection

For each researcher, we manually downloaded data from Web of Science (WoS) as the primary source, and used Scopus only when a WoS profile was unavailable. The raw data collected includes publication titles, citation counts, publication date, author information, and source details (e.g., publication venue, volume, issue, pages).

The dataset was collected manually by seven annotators, and the data of each researcher was subsequently reviewed and validated by the thesis author, who performed a comprehensive quality check. Initially, we considered using Google Scholar as the source of

bibliometric data. However, after careful consideration, we opt for the WoS and Scopus databases because of their higher data quality. Our analysis showed that Google Scholar often includes unverified content across the internet, listing any document containing the author's name as a publication, leading to unreliable data.

2.3 Data preprocessing

We put together all the researcher data that are needed to calculate the features in one place, either from WoS and Scopus whichever is applicable to the researcher. Then, we created a new document that includes a subset of the existing columns (i.e., name, field, title, authors, citations), along with additional calculations for each publication (i.e., a flag for indicating whether the researcher is the first author, a flag for author inclusion, the total number of authors, and the researcher's position in the author list) that are needed to calculate the features. The features of the model are described in detail in Section 2.4.

To ensure reliability of the dataset, our analysis; i) removed duplicate publications, (ii) made sure the author names are correctly processed (e.g., accented characters, suffix and prefixes such as "Jr.", "De", "Al"), (iii) verified authorship by confirming that each listed publication included the researcher's name as an author, (iv) filtered professors who had less than five publications in the dataset, and (v) filtered publications with publication years earlier than 1900 to ensure accuracy. Table 2.1 presents a descriptive summary of the dataset, which consists of a total of 296,234 publications (42% Nobel, 58% average) by 2,331 researchers (27% Nobel, 73% average). The average group includes researchers from 117 universities around the world. Please see Appendix A for the list of the universities.

TABLE 2.1: Data description.

Field name	Positive Class (higher academic quality)	Negative Class (lower academic quality)
Physics	182	511
Chemistry	173	514
Physiology or Medicine	185	344
Economics	92	330
Total number of researchers	632	1,699

2.4 Features for academic quality quantification

Based on a thorough literature review, we identified 15 features to be used as input for both machine learning models to predict academic quality. These features include a set of new metrics proposed in this thesis and conventional bibliometric metrics to train and evaluate models for academic quality quantification. Table 2.2 presents a descriptive summary of the features.

TABLE 2.2: Features for academic quality quantification.

Feature	Description
Conventional metrics	
<i>Number of publications</i>	Total number of publications
<i>Total citations</i>	Total number of citations
<i>h-index</i>	Number of papers (h) with citations $\geq h$
<i>normalized h-index</i>	h -index divided by the square root of the total number of publications
<i>g-index</i>	Largest number of top g publications whose cumulative citations reach at least g^2
<i>a-index</i>	Average number of citations of h -core publications
<i>k-index</i>	Average citations per publication, divided by the ratio of citations in the h -tail to citations in the h -core
<i>i10-index</i>	Number of publications with at least 10 citations
<i>i100-index</i>	Number of publications with at least 100 citations
<i>i1000-index</i>	Number of publications with at least 1000 citations
<i>i10000-index</i>	Number of publications with at least 10000 citations
<i>cnp-index</i>	Average number of citations per publication
Proposed new metrics	
<i>cnp-index</i>	Average number of citations per author per publication
<i>max_fa_citations</i>	Maximum number of citations among publications where the researcher is the first author
<i>aap-index</i>	Average author position across all publications

The following are the descriptions of the features.

- *Number of publications* (Number of publications per person): This metric reflects the overall research productivity of the researcher, defined as

$$\text{Number of publications} := \text{Pub}_{\text{count}}. \quad (2.1)$$

- *Total citations* (Sum of citations of all publications per person): This metric indicates the total impact or influence of a researcher's work, defined as

$$\text{Total citations} := \text{Cit}_{\text{all}}. \quad (2.2)$$

- *h-index* (Number of papers (h) with a citation count $\geq h$): The h-index is the maximum number h for which the top h most-cited publications have each received at least h citations. This metric reflects the most impactful and productive part of a researcher's work based on the number citations received, defined as

$$h\text{-index} := \max \{h \in \mathbb{N} : \text{Cit}_h \geq h\}. \quad (2.3)$$

- *normalized h-index* (h-index divided by the square root of total number of publications) [33]: This metric adjusts h-index for the number of publications, helping to compare researchers with different publication volumes, defined as

$$\text{normalized } h\text{-index} := \frac{h\text{-index}}{\sqrt{\text{Pub}_{\text{count}}}}. \quad (2.4)$$

- *g-index* (Largest number of top g publications whose cumulative citations reach at least g^2): The g-index is the maximum number g such that top g most-cited publications have at least g^2 citations in total. Similar to the h-index, the g-index evaluates impact and productivity of a researcher, however, it gives more weight on impact. A few highly cited publications can result in a higher g-index than a larger number of averagely cited ones, defined as

$$g\text{-index} := \max \left\{ k \in \mathbb{N} : \sum_{i=1}^k \text{Cit}_i \geq k^2 \right\}. \quad (2.5)$$

- *a-index* (Average number of citations of h-core articles): Since h-index and g-index values are integers, many researchers, especially in small groups, can share the same value, making it difficult to distinguish their performance. Moreover, h-index effectively highlights highly impactful researchers, but it has limited ability to differentiate among average researchers. This metric addresses this limitation as a-index does not have to be an integer value and may increase as the researcher's

citation counts increase, even if the h-index remains unchanged. This makes a-index more effective at distinguishing researchers with the same h-index and g-index value. a-index is defined as

$$a\text{-index} := \frac{\sum_{i=1}^h Cit_i}{h}. \quad (2.6)$$

- *k-index* (Average number of citations per publication divided by the ratio of citations in the h-tail to citations in the h-core) [34]: H-tail consists of publications having at most h citations and is not captured by h-index. To avoid this loss, this metric considers the citations in the less cited h-tail, which are not a part of highly cited h-core, defined as

$$k\text{-index} := \frac{\left(\frac{Cit_{all}}{Pub_{count}}\right)}{\left(\frac{Cit_{h\text{-tail}}}{Cit_{h\text{-core}}}\right)}. \quad (2.7)$$

- *i10-index* (Number of publications with ≥ 10 citations): This metric highlights a researcher's ability to consistently produce impactful work, defined as

$$i10 := |\{p \in Pub \mid Cit_p \geq 10\}|. \quad (2.8)$$

- *i100-index* (Number of publications with ≥ 100 citations) [35]: Publications with over 100 citations are generally considered influential in their field and signify a strong contribution to scientific discourse. This metric highlights a researcher's ability to consistently produce highly impactful work, defined as

$$i100\text{-index} := |\{p \in Pub \mid Cit_p \geq 100\}|. \quad (2.9)$$

- *i1000-index* (Number of publications with ≥ 1000 citations) [35]: High citation counts are rare and typically reflect groundbreaking or field-defining work. This metric serves as a marker of exceptional academic influence and long-term relevance, defined as

$$i1000\text{-index} := |\{p \in Pub \mid Cit_p \geq 1000\}|. \quad (2.10)$$

- *i10000-index* (Number of publications with ≥ 10000 citations) [35]: Extremely high citation counts are rare and typically reflect groundbreaking or field-defining work. This metric also serves as a marker of exceptional academic influence and long-term relevance, defined as

$$\text{i10000-index} := |\{p \in \text{Pub} \mid \text{Cit}_p \geq 10000\}|. \quad (2.11)$$

- *cpp-index* (Citations per publication): This metric normalizes the total citation count by productivity, allowing for a quality-over-quantity assessment. It provides insight into the impact of each publication on average and defined as

$$\text{cpp-index} := \frac{\text{Cit}_{\text{all}}}{\text{Pub}_{\text{count}}} = \frac{\sum_{i=1}^N \text{Cit}_i}{N}. \quad (2.12)$$

- *cpppp-index* (Citations per author per publication): By accounting for co-authorship, this metric emphasizes individual contribution and helps distinguish between highly collaborative and independent high-impact work, defined as

$$\text{cpppp-index} := \frac{1}{N} \sum_{i=1}^N \frac{\text{Cit}_i}{a_i}. \quad (2.13)$$

- *max_fa_citations* (Maximum number of citations among publications where the researcher is the first author): The first author often plays a key role in the execution and production of research. This metric captures leadership in the production of research, defined as

$$\text{max_fa_citations} := \max_{i \in \mathcal{F}} \text{Cit}_i. \quad (2.14)$$

- *aap-index* (Average author position): The author's position is often used to infer the role in the publication. A low aap index implies consistently early author position (e.g., first or second) and may indicate sustained leadership or major contributions in publications. aap index is defined as

$$\text{aap-index} := \frac{1}{N} \sum_{i=1}^N a_i. \quad (2.15)$$

2.5 Model selection

We identify the academic quality assessment problem as a supervised learning task, where the AQI is predicted as a value between 0 and 1. A higher AQI value, closer to 1, indicates greater academic quality. The models are trained using data from Nobel laureates (the "positive class"), who represent high-quality academic profiles, and data from average academics (the "negative class"), who have relatively lower-quality academic profiles based on their university rankings.

To predict AQI and identify the most important features for assessing academic quality, we use Siamese NN and XGBoost [36] as the ensemble learning model. These models incorporate: (i) a class separation margin, where applicable, which penalizes small margins between the positive and negative classes to maximize the dissimilarity between classes; (ii) regularization to prevent overfitting; and (iii) hyperparameter tuning to optimize the model performance.

Model training: For each modeling approach (Siamese NN and XGBoost), five models are developed using data from the following subsets: (i) the full dataset (all fields); (ii) Physics; (iii) Chemistry; (iv) Physiology or Medicine, and (v) Economics. For each subset, the corresponding Siamese NN and XGBoost models are trained and tested using the same train-test splits to ensure a fair comparison of their performance.

Data are split into train (80%) and test sets (20%), with performance evaluated using 5-fold cross-validation. Once the models are trained, the resulting AQI values provide a measure of researcher performance and feature importance scores that distinguish successful researchers.

Model quality: To evaluate the performance of all AQI prediction models, we use the MQ score, a custom metric to minimize variance of the same class and maximize differentiation between positive and negative classes. Inspired by Fisher's linear discriminant analysis, the MQ score is defined as

$$\text{MQ} := \frac{(m_p - m_n)^2}{s_p^2 + s_n^2},$$

where:

- m_p and m_n are the mean AQI values for the positive and negative classes, respectively,
- s_p and s_n are the variances of AQI values within the positive and negative classes.

Software: All data preprocessing and analysis, and implementation of the Siamese NN and XGBoost models are performed using Python programming language. For software packages and libraries, we use `PyTorch` and `scikit-learn` to implement Siamese NN, `XGBClassifier` and `scikit-learn` to implement XGBoost, and `matplotlib` for visualizations.

Hardware: The training of the models was conducted on a system with an Apple M2 Pro with 12-core CPU, 19-core GPU, advanced Neural Engine, and 16 GB RAM, which enabled efficient performance. The complete training process took 4.3 hours for all Siamese NN models and 0.3 hours for all XGBoost models.

Chapter 3

Models for Academic Quality Quantification

In this thesis, we employ two machine learning techniques, Siamese NN and XGBoost, to quantify academic quality. Both models are well-suited for supervised learning tasks involving complex nonlinear relationships, offering potentially improved performance.

The objective of all models is to predict academic quality by minimizing the distances between the same-class instances and maximizing the separation between positive and negative classes.

To evaluate model performance, we use the MQ score, a custom metric that rewards a larger mean separation between classes while penalizing high variance within each class.

3.1 Siamese neural network model

In this section, we introduce Siamese NN to distinguish between positive and negative classes, thus indirectly quantifying academic quality.

3.1.1 Overview of Siamese neural networks

A Siamese NN is a machine learning model, which learns the similarity between input paired data by comparing them using a distance metric that aims to distinguish and

classify. Although there are many similarity approaches, a Siamese NN is considered the most effective choice [37] to compare complex data samples with features of varying types.

Siamese NN consists of two identical neural networks that process two input vectors in parallel. Both networks are feedforward perceptrons trained using error backpropagation. They share the same architecture and weights, and each side learns a hidden representation (or features) of its input vector. The outputs of these networks are compared using a similarity function (e.g., distance function). The final output of the Siamese NN reflects how similar the two input vectors are based on their learned representations. An illustration of the Siamese NN model is shown in Figure 3.1.

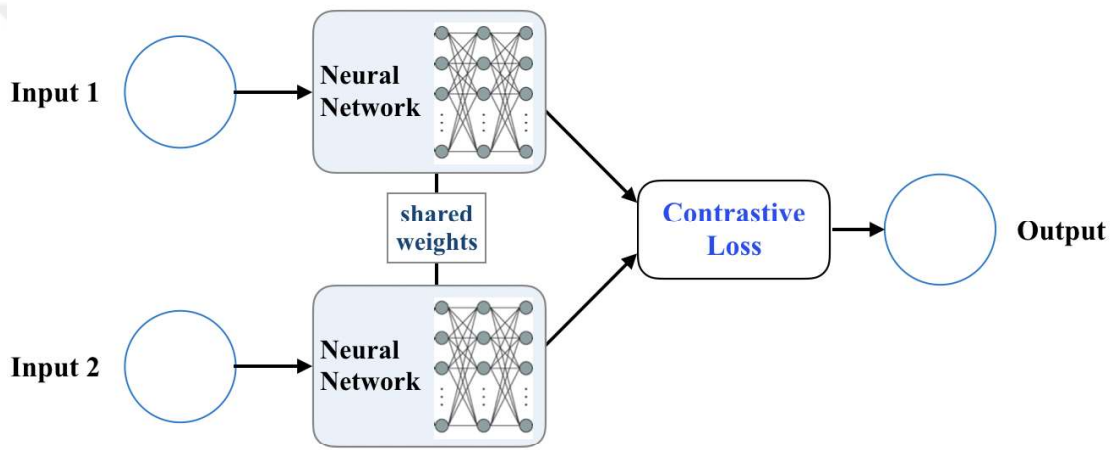


FIGURE 3.1: Siamese NN model.

Applications of Siamese neural network

Siamese NN was originally applied to signature verification to detect forged signatures by comparing two handwritten signatures [2]. It was able to state whether the two signatures were original or not. In recent years, research on Siamese networks has gained increased attention due to the improvement in computational capabilities [38]. They have since been widely used for classification problems by learning similarity metrics to compare and score input pairs, in several different fields such as image and video analysis, biology, chemistry, physics, medicine, signal processing, text mining, and robotics. For example, Siamese NN was used for action quality assessment [39], for medical imaging [40], for text similarity assessment [41], for chromosome classification [42], for road detection from the perspective of moving vehicles [43], for plant leaves classification [44], for software defect

prediction [45], for video anomaly prediction and detection [46], for phonetic similarity detection between speech signals [47] and for cervical cancer diagnosis [48].

Advantages and Challenges of Siamese neural networks

The main advantage of Siamese NN over other neural network models is their strong ability to learn effectively even with limited training data [49]. Accordingly, Siamese NN can effectively learn representations in limited or imbalanced data, or in high-dimensional settings, which are common challenges in binary classification. Table 3.1 presents a summary of the advantages and challenges of the Siamese NN.

TABLE 3.1: Advantages and challenges of Siamese NN.

Advantages of Siamese NN	Challenges of Siamese NN
Efficient learning from comparisons: Requires fewer labeled data by leveraging pairwise comparisons. Creating sample pairs augments the dataset size and allows the model to learn from pairwise similarity rather than absolute labels.	Computationally expensive: Generating all possible pairs leads to quadratic growth in training samples, increasing computation costs.
Flexibility in input data types: Can be used for tabular data, images, text, or mixed types.	Hyperparameter sensitivity: Requires careful tuning of margins used in contrastive loss function and batch sizes.
Better generalization: Siamese networks can generalize to unseen data via learned similarity metrics, especially in few-shot learning.	Training instability: Sensitive to initializations and learning rates due to pair-based optimization.

3.1.2 Siamese neural network for academic quality quantification

We implement the Siamese NN models to predict academic quality as follows. Section 3.1.2.1 explains the structure of the Siamese NN that learns the tabular data. Section 3.1.2.2 presents the modified contrastive loss function used to group same-class samples, to push different-class samples apart and to ensure that positive samples rank higher than negative ones. Section 3.1.2.3 explains the grid search performed over margin hyperparameters and number of epochs to find optimal settings. Finally, Section 3.1.2.4 shows the training of the final model with the best parameters and the testing process.

3.1.2.1 Siamese neural network architecture

We implement a Siamese NN model to predict academic quality by comparing pairs of researcher profiles from the dataset with 15 features and binary labels (positive class labeled as 1 and negative class labeled as 0). The model architecture is a fully connected feedforward neural network (multi-layer perceptron) with:

- Input layer: Input dimension is equal to the number of features (i.e., 15 features) in the dataset;
- 1st hidden layer: 48 units, ReLU activation, followed by 30% dropout for regularization;
- 2nd hidden layer: 48 units, ReLU;
- 3rd hidden layer: 48 units, ReLU;
- Output layer: 1 unit. The final output is a sigmoid activation to produce a score between 0 and 1 (interpreted as AQI).

This structure with regularization (via dropout) is well suited for our structured tabular data and data size, where a fixed-length feature vector is compared in pairwise fashion. An illustration of the Siamese NN architecture used for the prediction of AQI is shown in Figure 3.2.

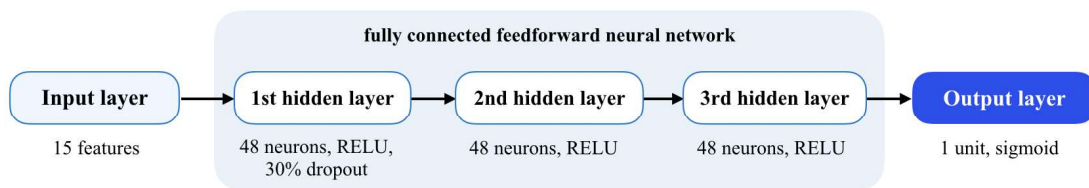


FIGURE 3.2: Siamese NN for AQI quantification.

Data preparation: Before training, five datasets were prepared: one comprising the full dataset (all fields), and four field-specific subsets, Physics, Chemistry, Physiology or Medicine and Economics, partitioned from the full dataset. Each dataset was split into 80% train and 20% test sets. Standard scaling (zero mean, unit variance normalization) was applied to all features. The scaler is fitted using the training data and subsequently applied to both the training and test sets.

Subsequently, all pairs consisting of positive and negative samples were generated for each model using full cross-pairing to construct the training inputs. Given P positive instances and N negative instances, we form $P \times N$ training pairs.

- Full dataset: $632 \times 1699 \times 80\% = 859,014$ pairs
- Physics: $182 \times 511 \times 80\% = 74,402$ pairs
- Chemistry: $173 \times 514 \times 80\% = 50,912$ pairs
- Physiology or Medicine: $185 \times 344 \times 80\% = 63,640$ pairs
- Economics: $92 \times 330 \times 80\% = 24,288$ pairs

Training: During training, each input instance x_i and x_j (i.e., input vectors of two researcher profiles) is independently passed through an identical feedforward neural network f_θ with shared weights θ to produce comparable embeddings

$$h_i = f_\theta(x_i),$$

$$h_j = f_\theta(x_j),$$

where:

- $f_\theta(\cdot)$ is the neural network with parameters θ ,
- h_i, h_j are the output embeddings (feature representations) of the two profiles.

At each hidden layer l , the neural network performs a non-linear activation

$$a^{(l)} = \phi \left(W^{(l)} a^{(l-1)} + b^{(l)} \right),$$

where:

- $a^{(0)} = x$ is the input vector,
- $W^{(l)}$ and $b^{(l)}$ are the weights and biases at layer l ,
- $\phi(\cdot)$ is the ReLU activation function,
- $a^{(l)}$ is the output activation of layer l .

These activations transform the inputs into high-level representations, enabling the network to capture non-linear relationships between features. The embeddings h_i and h_j are the outputs of the final layer.

Weight updates via backpropagation: The loss function $\mathcal{L}$ (defined in Section 3.1.2.2) compares these outputs of sample pairs to quantify how well the model distinguishes between positive and negative samples. $\mathcal{L}$ encourages embeddings of similar instances to be close and dissimilar instances to be far apart. The network parameters are updated by minimizing $\mathcal{L}$ using stochastic gradient descent (SDG) as follows:

1. Compute gradients of the loss with respect to sigmoid-transformed outputs as

$$\frac{\partial \mathcal{L}}{\partial s_i}, \quad \frac{\partial \mathcal{L}}{\partial s_j}.$$

2. Backpropagate through the sigmoid and each layer to compute gradients for all layers l as

$$\frac{\partial \mathcal{L}}{\partial W^{(l)}}, \quad \frac{\partial \mathcal{L}}{\partial b^{(l)}}.$$

3. Update weights and biases with learning rate η as

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial W^{(l)}}, \quad b^{(l)} \leftarrow b^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial b^{(l)}}$$

Through these iterative updates, the neural network learns to produce embeddings that bring similar instances closer and push dissimilar instances apart, improving the model's ability to predict academic quality.

Similarity score computation. The Siamese NN outputs the scores that indicate the similarity between two input instances. These scores are computed using a sigmoid function

$$\sigma(h_i) = \frac{1}{1 + e^{-h_i}},$$

$$\sigma(h_j) = \frac{1}{1 + e^{-h_j}},$$

where $\sigma(h_i)$ and $\sigma(h_j)$ are the predicted scores (AQI value) between 0 and 1 for each input instance (i.e., researcher profile). These scores are then used in the loss function to guide the training process.

Optimization Setup. The model training is performed using stochastic gradient descent with a learning rate of 1×10^{-3} and batches of size 32. To reduce overfitting and enhance generalization, we apply the following:

- Early stopping (patience = 10 epochs),
- Dropout regularization,
- Lower model complexity (48 neurons),
- 5-fold cross-validation.

3.1.2.2 Loss function: Modified contrastive loss

In a standard Siamese NN, the input to the model is a pair of feature vectors that are passed through the same neural network to generate their respective embeddings. The similarity or dissimilarity between these embeddings is then quantified, commonly through contrastive or triplet loss functions.

In this thesis, the Siamese NN model is trained using a modified contrastive loss to maximize separation between embeddings of different classes, while preserving same-class similarity. Modified contrastive loss function, $\mathcal{L}_{\text{mcl}}$, is defined as

$$\mathcal{L}_{\text{mcl}} := \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{order}}, \quad (3.1)$$

$$\mathcal{L}_{\text{contrastive}} = \text{Var}(f_P) + \text{Var}(f_N) + \frac{1}{N} \sum_{i=1}^N \left[\max \left(0, m_c - \left| f_P^{(i)} - f_N^{(i)} \right| \right) \right]^2, \quad (3.2a)$$

$$\mathcal{L}_{\text{order}} = \frac{1}{N} \sum_{i=1}^N \max \left(0, m_o - f_P^{(i)} + f_N^{(i)} \right), \quad (3.2b)$$

where

- N is the number of pairs
- f_P are the output values for positive class
- f_N are the output values for negative class

- $f_P^{(i)}$ is the output value for the i -th positive sample
- $f_N^{(i)}$ is the output value for the i -th negative sample
- m_c is the contrastive loss margin
- m_o is the order loss margin

L_{mcl} comprises two distinct components: contrastive loss $L_{\text{contrastive}}$ and order loss L_{order} .

- **Contrastive loss term** ($L_{\text{contrastive}}$): This term penalizes when the outputs of positive and negative pairs are not sufficiently separated and enforces compact clusters by minimizing the same-class output variance as follows:
 - Different-class separation: For pairs (x_i, x_j) of different classes (i.e., one positive and one negative), their outputs are separated by at least the specified margin m_c .
 - Same-class clusters: Minimizes the variance of outputs among same-class pairs (both positive and negative), encouraging them to cluster closely.
- **Order loss term** (L_{order}): This term enforces that positive samples receive higher scores than negative ones by at least the margin m_o , as samples in the positive class are expected to exhibit higher academic quality compared to those in the negative class.

3.1.2.3 Hyperparameters of Siamese neural network

To train the models effectively, we perform a hyperparameter grid search for the margin parameters using a 20% validation set, and a separate grid search for the optimal number of epochs for early stopping, using a 5-fold cross-validation.

Margin parameters: For the model trained on the full dataset (all fields), a grid search was performed to find the optimal margin values for the modified contrastive loss function. The following margin parameters are tested using the 20% validation set, which was split from the initial 80% training set (resulting in a 60% training, 20% validation, and 20% test split) $m_c \in \{0.25, 0.5, 0.75\}$, and $m_o \in \{0.2, 0.3, 0.4, 0.5\}$.

Each model was trained with the same architecture and optimizer (SGD, lr=0.001) and evaluated by the MQ score. The optimal margin parameters obtained from the full dataset were subsequently used to train the four field-specific Siamese NN models as well.

Early stopping: For each of the five Siamese NN models, an additional grid search was performed using a 5-fold cross-validation to determine the optimal number of training epochs (up to a maximum of 100 epochs) using early stopping. The best number of training epochs is decided based on the fold with the highest average MQ score. Table 3.2 presents the final training parameters.

TABLE 3.2: Training parameters for the Siamese NN model.

Training Parameters					
Parameter	Physics	Chemistry	Physiology or Medicine	Economics	All Fields
m_c	0.75	0.75	0.75	0.75	0.75
m_o	0.2	0.2	0.2	0.2	0.2
Number of epochs	88	66	74	100	40
Learning rate	0.001	0.001	0.001	0.001	0.001
Batch size	32	32	32	32	32

3.1.2.4 Final model training and testing

Final Siamese NN models were trained using:

- The training datasets with all pairs from positive and negative classes;
- The same architecture;
- The best margins and number of epochs identified during grid search.

The outcome is a trained Siamese neural network that maps the input feature vectors (from tabular data) into a similarity score (between 0 and 1) such that:

- Both positive class (label = 1) and negative class (label = 0) samples are clustered in the output space (close same-class scores).
- Different class samples are separated (by at least contrastive loss margin m_c).
- Positive samples score higher than negative samples (by at least order loss margin m_o).

The models output a scalar between 0 and 1 for each test sample that can be interpreted as AQI.

3.2 Ensemble learning model

In this section, we introduce XGBoost, a powerful gradient boosting model, to quantify academic quality.

3.2.1 Overview of XGBoost

XGBoost was first introduced in 2016 [36], and is considered one of the most effective gradient boosting implementations. It constructs an ensemble of decision trees, where each tree is trained to fix the errors (residuals) of the previous trees. This is achieved by minimizing a regularized loss function using gradient descent.

Specifically, at each boosting step, the newly trained tree predicts the negative gradients (residuals) of the loss function with respect to the model's output. Its output is then added to the prediction of the previous trees to improve the final output of the model. The boosting process continues iteratively until either the residuals stop reducing or the maximum number of trees is reached.

One of the valuable aspects of XGBoost is its built-in feature importance metrics, which help to understand how much each feature influences the model's predictions.

Applications of XGBoost

Due to its robust predictive performance and relatively low computational cost, XGBoost has been adopted across various domains, ranging from chronic kidney disease diagnosis [50], bankruptcy prediction [51], and credit risk assessment [52] to vehicle driving assessment [53]. There is also research [54] that demonstrates that XGBoost can perform better than other machine learning methods on imbalanced datasets.

Advantages and challenges of XGBoost

The main advantage of XGBoost is its strong learning ability with tabular data and its efficiency in managing limited or imbalanced datasets, which makes it well-suited

model for our study, where the class imbalance is 2.7:1 (1699 negative samples vs. 632 positive samples). Table 3.3 presents a summary of the advantages and challenges of the XGBoost.

TABLE 3.3: Advantages and challenges of XGBoost.

Advantages of XGBoost	Challenges of XGBoost
High predictive performance: Utilizes gradient boosting and regularization (L1/L2), making it highly accurate and less likely to overfit compared to basic decision trees.	Hyperparameter complexity: Requires many hyperparameters, which can be computationally intensive.
Robust with structured/tabular data: XGBoost often outperforms neural networks on small-to-medium structured datasets.	Less effective for unstructured data: Struggles with image, audio, or text data compared to neural networks.
Efficient and scalable: Supports parallelized tree building and out-of-core computation.	Overfitting risk on noisy data: Although regularized, it may still overfit if the dataset contains too much noise or redundant features.

3.2.2 XGBoost for academic quality quantification

We implement XGBoost models to predict academic quality as follows. Section 3.2.2.1 explains the structure of the XGBoost that learns the tabular data. Section 3.2.2.2 presents the loss function. Section 3.2.2.3 explains the grid search performed over hyperparameters to find optimal settings. Finally, Section 3.2.2.4 shows the training of the final model with the best parameters and the testing process.

3.2.2.1 XGBoost architecture

We implement XGBoost model to predict the probability that an instance belongs to the positive class using the dataset with 15 features and binary labels (positive class labeled as 1 and negative class labeled as 0). XGBoost first predicts the *log-odds* (logits) that a researcher belongs to the positive class. These logits are then converted into probabilities using the sigmoid function. The resulting probabilities are interpreted as a measure of the AQI for each researcher.

- A high probability indicates that the researcher belongs to the positive class, implying higher academic quality.

- A low probability indicates that the researcher belongs to the negative class, implying lower academic quality.

Data preparation: We use the same train (80%) and test (20%) splits prepared for the Siamese NN models to train and evaluate XGBoost models: one comprising the full dataset (all fields), and four field-specific subsets, Physics, Chemistry, Physiology or Medicine and Economics, partitioned from the full dataset. The dataset is not standardized for XGBoost, as ensemble learning models do not require feature scaling.

Initialization: XGBoost initializes the prediction y_0 for every instance x_i in the dataset as

$$y_0 = 0.5 \quad \text{as an initial guess, which will be iteratively improved during training.} \quad (3.3)$$

Training - iterative boosting process: XGBoost implements the iterative boosting process, in which an ensemble of decision trees is built to correct the errors (residuals) of previous trees as follows:

1. Compute residuals: For each instance, the residual r_i is the difference between the true label y_i and the current prediction $\hat{y}_i$, defined as

$$r_i = y_i - \hat{y}_i \quad \text{for each training instance } (i),$$

where y_i is the true label (0 or 1) and $\hat{y}_i$ is the model's prediction (initially 0.5).

2. Build a decision tree: At each iteration, XGBoost fits a new decision tree to predict the residuals. The tree is trained to minimize the logistic loss function (in the case for binary classification), effectively learning the necessary corrections to improve the model's predictions.

3. Calculate leaf values for the tree: The leaf value indicates the prediction of the instances within that leaf node and is used to update the model's overall prediction. It is calculated as the negative average of the residuals for all instances that fall into current leaf node as

$$\text{leaf value} = \frac{-\sum r_i}{\text{number of instances in the leaf}}.$$

After a new tree decision tree is built, its output is added to the cumulative predictions of the existing model (previous trees) as

$$\hat{y}_i^{\text{new}} = \hat{y}_i^{\text{old}} + \text{learning rate} \times \text{leaf value},$$

where:

- $\hat{y}_i^{\text{old}}$ is the previous prediction before adding the new tree.
 - The leaf value is the output (prediction) of the current leaf node in this new tree, calculated from residuals.
 - The learning rate controls the magnitude of the update; a smaller learning rate means that each tree makes smaller corrections.
 - $\text{learning rate} \times \text{leaf value}$ is the impact of the new tree.
 - $\hat{y}_i^{\text{new}}$ is the updated prediction after including this tree.
4. Repeat iterative boosting to build an ensemble of trees: The boosting process repeats until the residuals stop improving or the maximum number of trees is reached. Each new tree corrects the errors of the previous trees by adjusting the predictions, improving the accuracy of the model over time.

Final model prediction: Once the trees are trained, XGBoost computes a logit score for each instance x_i , which is the sum of the predictions from all trees scaled by the learning rate. The predicted output (logit score) for an instance x_i is defined as

$$\hat{y}_i^{\text{final}} = \hat{y}_i^0 + \sum_{t=1}^T f_t(x_i),$$

where $f_t(x_i) = \text{learning rate} \times \text{leaf value}_t$

Here, $f_t(x_i)$ is the contribution (prediction) from the t -th tree for instance x_i , and T is the number of trees in the model.

An illustration of the XGBoost model used for the academic quality quantification is shown in Figure 3.3.

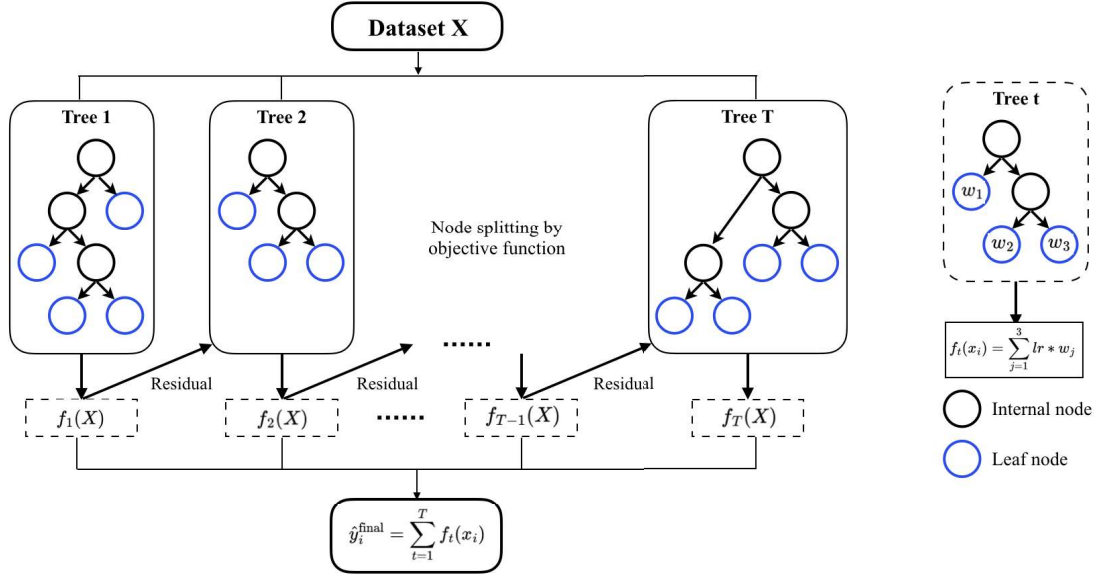


FIGURE 3.3: XGBoost model for AQI quantification (adapted from [1]).

Conversion to probability (using logistic function): The logit score is then passed through a logistic (sigmoid) function to convert it into a probability that the instance belongs to the positive class as

$$P(y_i = 1 | x_i) = \frac{1}{1 + e^{-\text{logit}(x_i)}},$$

where:

- $P(y_i = 1 | x_i)$ is the predicted probability of the model that instance i belongs to class 1 (positive class).
- $\text{logit}(x_i)$ is the output from the ensemble of decision trees.

This probability is interpreted as an indicator of how likely an instance (i.e., researcher) belongs to the positive class. For example, let us say we have a researcher with a set of 15 feature values (e.g., h-index, total citations, etc.). After passing the researcher's features through the trained XGBoost model, the predicted probability of the researcher is

$$P(y_i = 1 | x_i) = 0.75.$$

This indicates that the researcher's AQI value is 0.75. The closer the AQI is to 1, the higher the academic quality.

3.2.2.2 Loss function

The model minimizes classification error by updating the trees using the *gradients* (first derivatives) of the loss function with respect to the model predictions. To differentiate between the two classes, XGBoost employs a *logistic loss function*, $l_{\log}$, defined as

$$l_{\log}(y_i, \hat{y}_i) := -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \quad (3.4)$$

where $y_i \in \{0, 1\}$ is the true class label and $\hat{y}_i$ is the predicted probability for sample i . During training, we minimize the following function

$$\mathcal{L}_{\text{ensemb}} := \sum_{i=1}^N l_{\log}(y_i, \hat{y}_i) + \Omega(f), \quad (3.5)$$

where $\Omega(f)$ is the regularization term that controls model complexity, based on the structure of the decision trees. XGBoost incorporates both L1 (lasso) and L2 (ridge) regularization within $\Omega(f)$ to control overfitting and improve generalization.

3.2.2.3 Hyperparameters of XGBoost

To train the models effectively, we performed a hyperparameter grid search using 5-fold stratified cross-validation. The parameters of grid search are:

- **n_estimators** $\in [30, 35, 40, 45, 50]$. The number of weak learners (decision trees).
- **max_depth** $\in [3, 5, 7]$. The maximum depth allowed for each of the weak learners. The model grows the tree up to the *max_depth* and then performs backward pruning, removing nodes until the loss falls below a threshold. Controls the complexity of individual trees, where shallower trees help to reduce overfitting.
- **learning_rate** $\in [0.01, 0.05]$. Step size parameter that helps to reduce overfitting, where lower values lead to slower, more cautious convergence.
- **subsample** $\in [0.8, 1]$. The proportion of the training data randomly sampled to grow each tree, where values below 1 introduce randomness and reduce overfitting.

- `min_child_weight` $\in [3, 5, 7, 10]$. The minimum total weight of samples needed to form a child node. The higher value restricts splitting nodes with few samples, which helps to reduce overfitting.
- `reg_lambda` $\in [1, 5, 10]$. L2 regularization term, which is useful when features are collinear.
- `reg_alpha` $\in [0.01, 0.1]$. L1 regularization term that promotes feature selection.
- `scale_pos_weight` $\in [1, 2, 2.7, 2.8, 3, 3.6, 4]$. This parameter is used to compensate for class imbalance. Balances the importance of positive versus negative classes by scaling the loss associated with positive class instances to allow the model to pay more attention to this underrepresented class. For example, 2.7 is calculated as the ratio of negative to positive samples (1699:632), to reflect the class distribution.
- `gamma` $\in [0, 0.1, 0.5]$. The minimum decrease in loss (or gain) necessary for a node to be split. This parameter regularizes tree growth by controlling node splits and reduces overfitting for small and imbalanced data. A node is split only if the resulting split leads to a loss reduction greater than gamma. Higher gamma implies a more conservative tree with fewer splits.

Each model was evaluated according to the MQ score to select the best parameters. Table 3.4 presents the final training parameters. This configuration ensures that the

TABLE 3.4: Training parameters for the XGBoost model.

Training Parameters					
Parameter	Physics	Chemistry	Physiology or Medicine	Economics	All Fields
<code>n_estimators</code>	50	50	50	30	50
<code>max_depth</code>	5	5	5	3	5
<code>learning_rate</code>	0.05	0.05	0.05	0.05	0.05
<code>subsample</code>	1	1	1	0.8	1
<code>min_child_weight</code>	5	5	3	5	3
<code>reg_lambda</code>	1	1	1	1	1
<code>reg_alpha</code>	0.1	0.01	0.01	0.01	0.1
<code>gamma</code>	0.5	0.1	0.3	0	0.1
<code>scale_pos_weight</code>	2.8	3.0	4.0	3.6	2.7

XGBoost models are leveraging regularization, class balancing, and conservative tree structures to minimize overfitting.

3.2.2.4 Final model training and testing

Final XGBoost models were trained using:

- The same training datasets with Siamese NN models;
- The optimal hyperparameters found during grid search.

The outcome is a trained XGBoost model that maps the input feature vectors (from tabular data) to a probability (between 0 and 1) for each test sample that can be interpreted as AQI.

3.3 Feature importance determination

Feature importance scores quantify how much each input feature contributes to the model's predictions. Features with higher scores are more important for the model's output.

This is different from feature selection, which aims to identify only a subset of features that improve model performance. While feature importance scores provide information on the relevance of all features after training, feature selection uses such insights to exclude less useful features before or during model training. Feature importance helps interpreting the model but does not reduce features. Table 3.5 presents a comparison of the two approaches.

TABLE 3.5: Feature importance determination vs. Feature selection.

	Feature Importance (as applied in this study)	Feature Selection
Purpose	Quantifies importance of each feature to the model's prediction.	Selects a subset of features for building a more interpretable model.
Output	Importance scores for all features.	A subset of features.
When is it applied	After training, for interpretation.	Before or during training to reduce dimensionality.
How is it applied	Measure effect of perturbing each feature on model output.	Methods include filter (e.g., correlation) and embedded (e.g., Lasso).

3.3.1 Feature importance for Siamese neural network

In Siamese NN, the perturbation approach is used to determine the feature importance scores as follows:

- Compute the final model's predictions on the training data.
- Apply perturbation. For each feature $x^{(j)}$, create a copy of input tensor and zero out feature $x^{(j)}$ across all samples. This simulates "removing" feature $x^{(j)}$ by neutralizing its influence.
- Run the model on the perturbed data and compute predictions.
- Calculate the absolute difference between the original and perturbed output (predictions) per sample and take the average difference over all samples. This difference measures the change in the model output when the feature $x^{(j)}$ is removed and indicates how important that feature is. If setting the feature $x^{(j)}$ to zero causes a significant change in the model's output, the feature is considered more crucial; whereas if zeroing it out barely affects output, the model does not rely heavily on the feature.
- Normalize all importance scores so that they sum to 1.

This process is illustrated below. Let:

- $x_n = [x_n^{(1)}, x_n^{(2)}, \dots, x_n^{(15)}]$ be the n -th input data point, where each $x_n^{(j)}$ is a feature value.
- $f(x_n)$ be the model output for input x_n .
- $x_n^{\setminus j}$ be identical to x_n except that the feature $x^{(j)}$ is set to zero as

$$x_n^{\setminus j} := [x_n^{(1)}, \dots, x_n^{(j-1)}, 0, x_n^{(j+1)}, \dots, x_n^{(15)}].$$

The raw importance score of feature $x^{(j)}$ is calculated as

$$I(x^{(j)}) := \frac{1}{N} \sum_{n=1}^N |f(x_n) - f(x_n^{\setminus j})|.$$

To express feature $x^{(j)}$'s contribution as a percentage over all features, we normalize

$$\text{FI}(x^{(j)}) := \frac{I(x^{(j)})}{\sum_{k=1}^{15} I(x^{(k)})} \times 100.$$

3.3.2 Feature importance for ensemble learning

In XGBoost, feature importance is evaluated using gain-based metrics. Accordingly, each feature's contribution to predictive performance is measured by calculating the reduction in the logistic loss function when a node in the XGBoost decision trees is split using that feature. Features with higher gain values are considered to be more influential in reducing classification error. For a given feature $x^{(j)}$, its total gain is

$$\text{Gain}(x^{(j)}) = \sum_{\text{splits on } x^{(j)}} \text{Gain}_{\text{split}}. \quad (3.6)$$

Each split's gain is:

$$\text{Gain}_{\text{split}} := \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma, \quad (3.7)$$

where:

- G_L, G_R are the gradient sums (first derivatives of the loss) for the left and right child nodes as

$$G_L = \sum_{i \in \text{left node}} \frac{\partial \ell(y_i, f(x_i))}{\partial f(x_i)}, \quad (3.8)$$

$$G_R = \sum_{i \in \text{right node}} \frac{\partial \ell(y_i, f(x_i))}{\partial f(x_i)}. \quad (3.9)$$

- H_L, H_R are the Hessian sums (second derivatives of the loss) for the left and right child nodes as

$$H_L = \sum_{i \in \text{left node}} \frac{\partial^2 \ell(y_i, f(x_i))}{\partial f(x_i)^2}, \quad (3.10)$$

$$H_R = \sum_{i \in \text{right node}} \frac{\partial^2 \ell(y_i, f(x_i))}{\partial f(x_i)^2}. \quad (3.11)$$

- $\ell(y_i, f(x_i))$ is the logistic loss function,
- $f(x_i)$ is the model prediction (logit or score) for the i -th sample,
- λ is the L2 regularization parameter,
- γ is the regularization on the number of splits.

The feature importance score for feature $x^{(j)}$, denoted by $\text{FI}(x^{(j)})$, measures how useful each feature is in improving the model prediction, and is defined as

$$\text{FI}(x^{(j)}) := \frac{1}{S(x^{(j)})} \sum_{s=1}^{S(x^{(j)})} \text{Gain at split } s, \quad (3.12)$$

where $S(x^{(j)})$ counts how many times the feature $x^{(j)}$ is selected to split throughout the entire ensemble of trees.



Chapter 4

Results

4.1 Results of all-field data

The training and test performance of the Siamese NN model developed for AQI using the full dataset (all fields) are presented in the following figures. Figure 4.1 shows the training predictions and the prediction distribution, Figure 4.2 shows the test predictions and the prediction distribution and Figure 4.3 displays the training loss and MQ score per epoch.

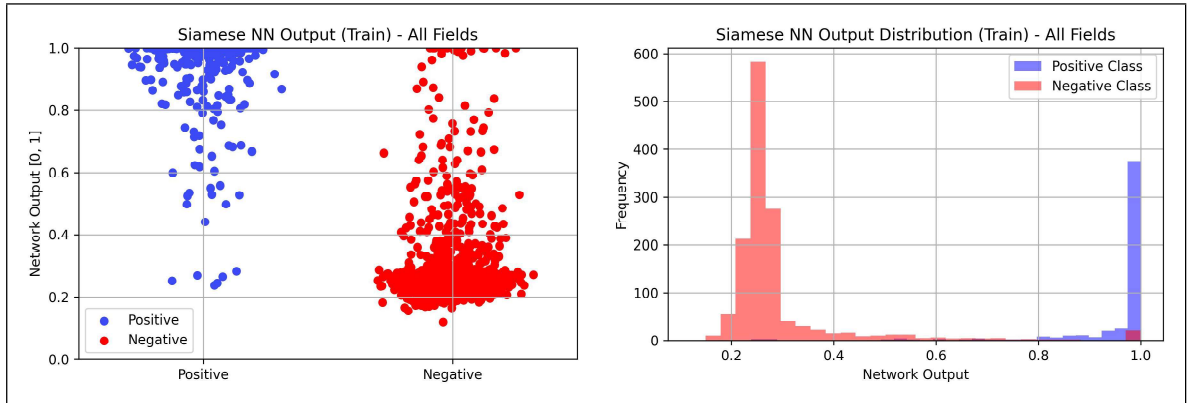


FIGURE 4.1: AQI distribution for Siamese NN model (training) - All Fields

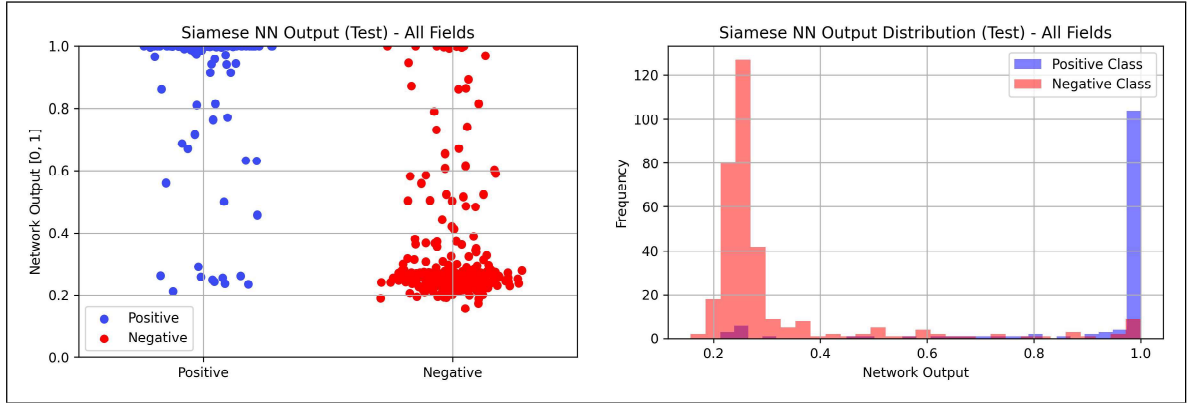


FIGURE 4.2: AQI distribution for Siamese NN model (test) - All Fields

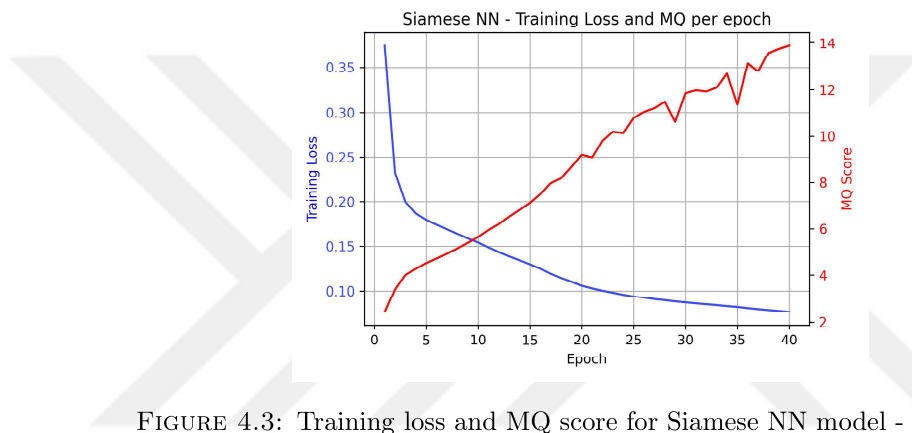


FIGURE 4.3: Training loss and MQ score for Siamese NN model - All Fields

The training and test performance of the XGBoost model developed for AQI using the full dataset (all fields) are presented in the following figures. Figure 4.4 shows the training predictions and the prediction distribution, and 4.5 shows the test predictions and the prediction distribution.

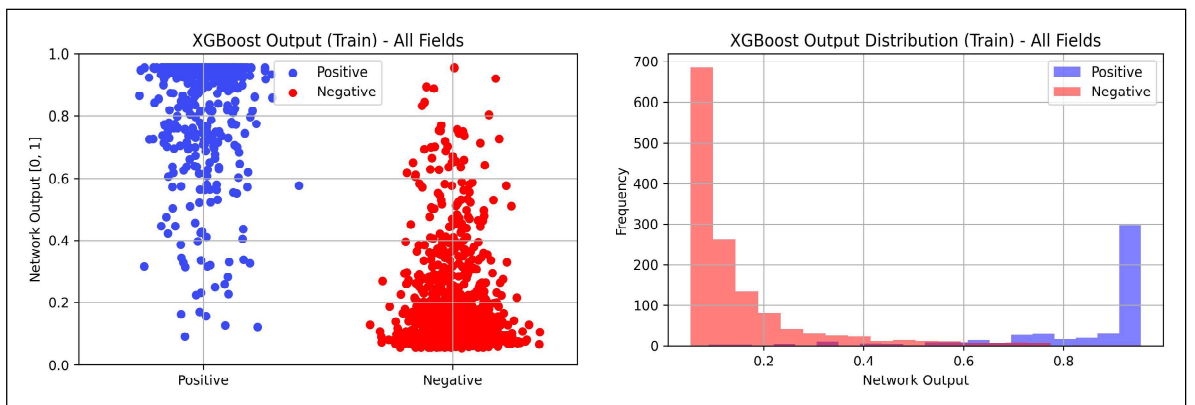


FIGURE 4.4: AQI distribution for XGBoost model (training) - All Fields

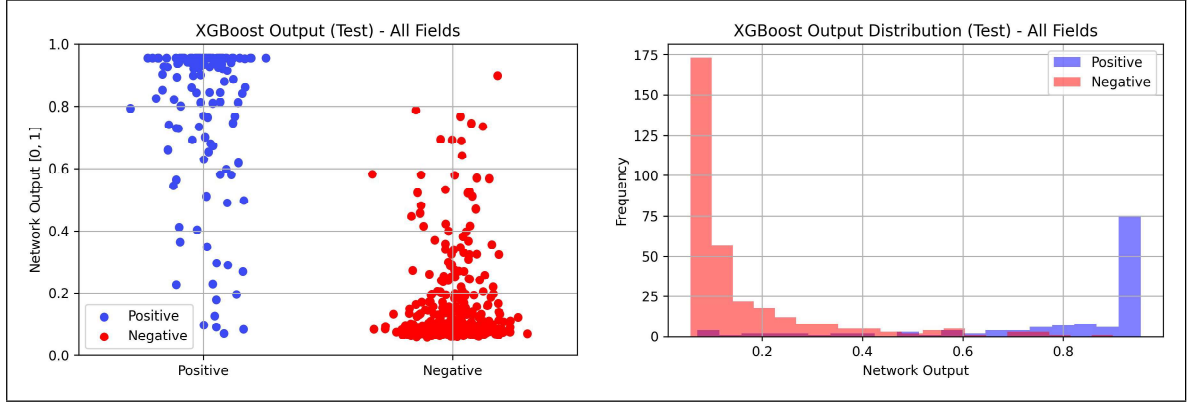


FIGURE 4.5: AQI distribution for XGBoost model (test) - All Fields

4.2 Results of field-based data

Integrating the dataset across different fields may have limitations, since each field possesses unique traits, and features important to one field may not be applicable to another. To address this, we developed field-specific models to quantify AQI and feature importance scores.

4.2.1 Physics

The training and test performance of the Siamese NN model developed for AQI using Physics data are presented in the following figures. Figure 4.6 shows the training predictions and the prediction distribution, Figure 4.7 shows the test predictions and the prediction distribution, and Figure 4.8 displays the training loss and MQ score per epoch.

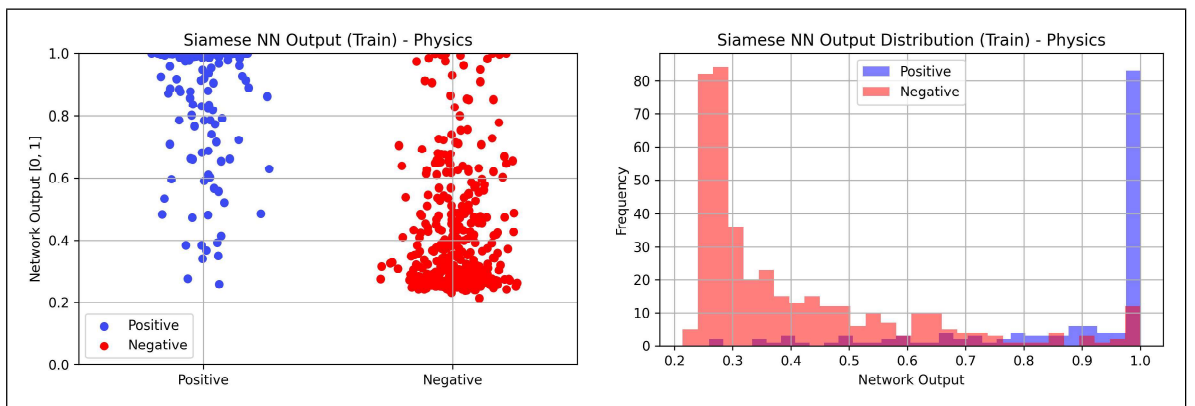


FIGURE 4.6: AQI distribution for Siamese NN model (training) - Physics

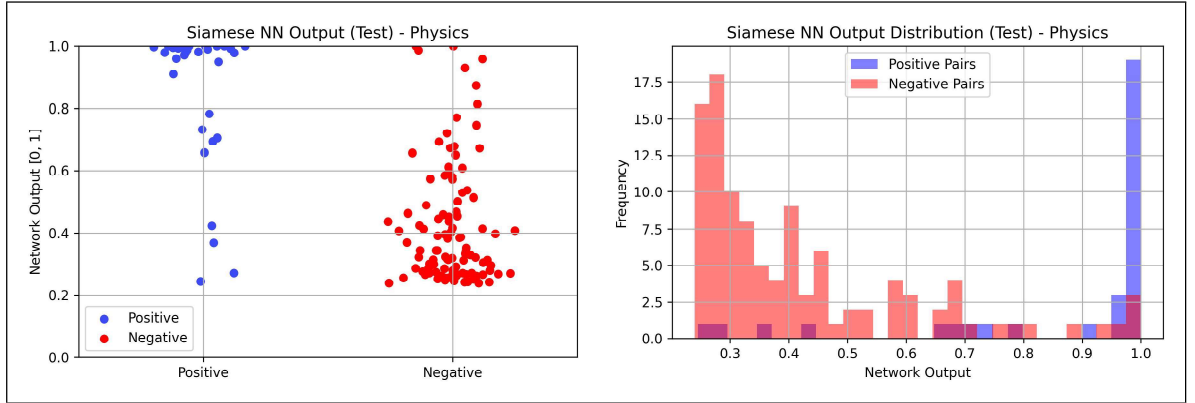


FIGURE 4.7: AQI distribution for Siamese NN model (test) - Physics

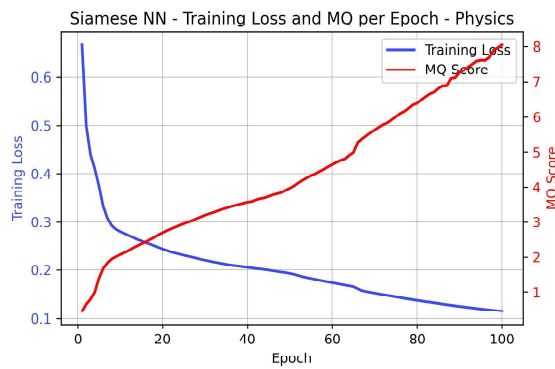


FIGURE 4.8: Training loss and MQ score for Siamese NN model - Physics

The training and test performance of the XGBoost model developed for AQI using Physics data is presented in the following figures. Figure 4.9 shows the training predictions and the prediction distribution, and Figure 4.10 shows the test predictions and the prediction distribution.

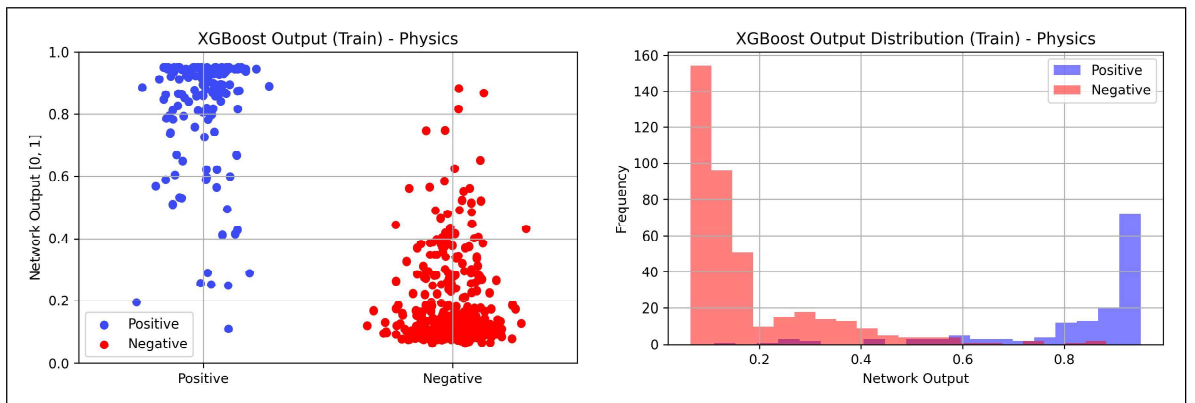


FIGURE 4.9: AQI distribution for XGBoost model (training) - Physics

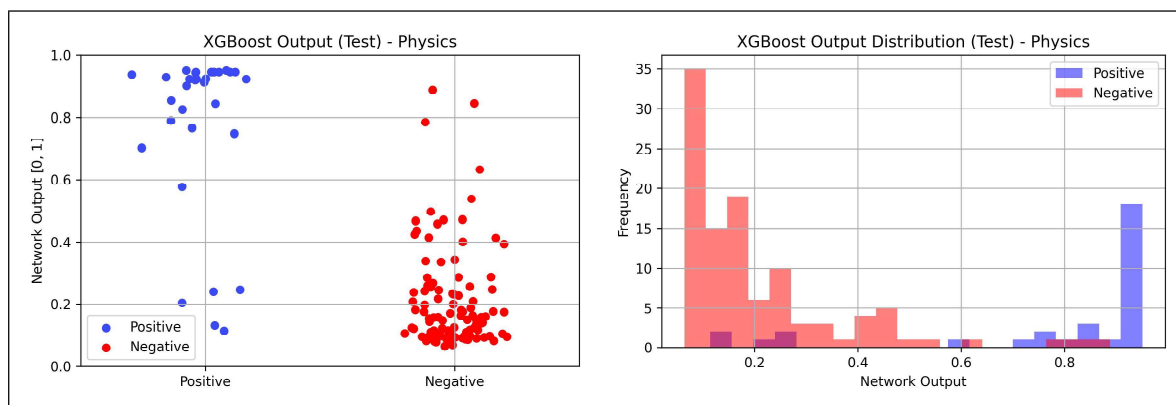


FIGURE 4.10: AQI distribution for XGBoost model (test) - Physics

4.2.2 Chemistry

The training and test performance of the Siamese NN model developed for AQI using Chemistry data are presented in the following figures. Figure 4.11 shows the training predictions and the prediction distribution, Figure 4.12 shows the test predictions and the prediction distribution, and Figure 4.13 displays the training loss and MQ score per epoch.

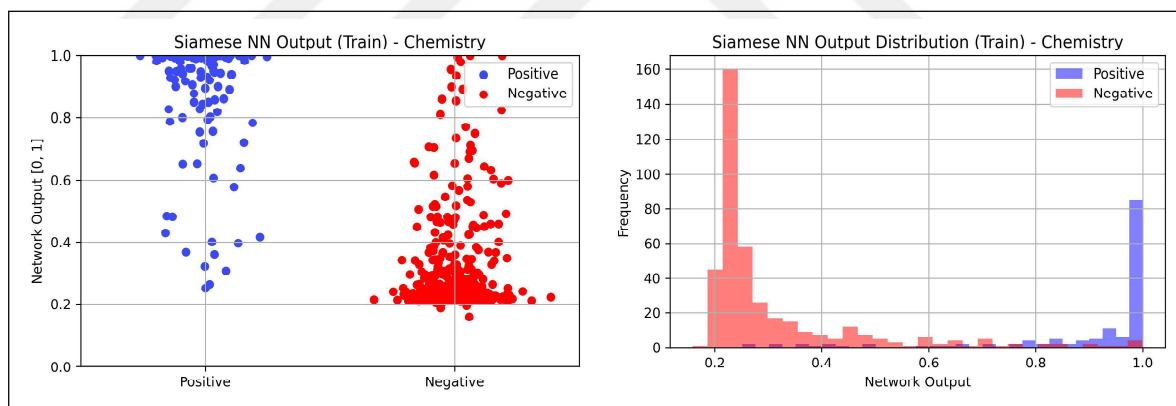


FIGURE 4.11: AQI distribution for Siamese NN model (training) - Chemistry

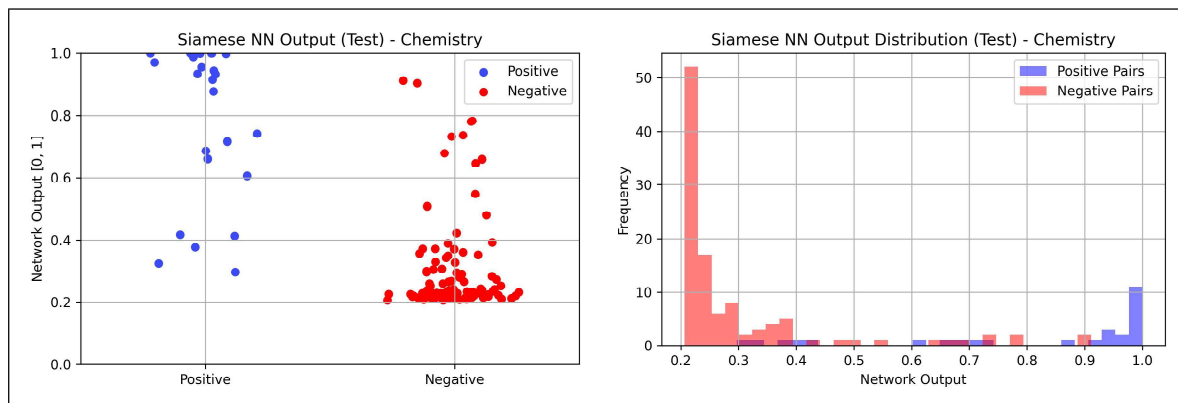


FIGURE 4.12: AQI distribution for Siamese NN model (test) - Chemistry

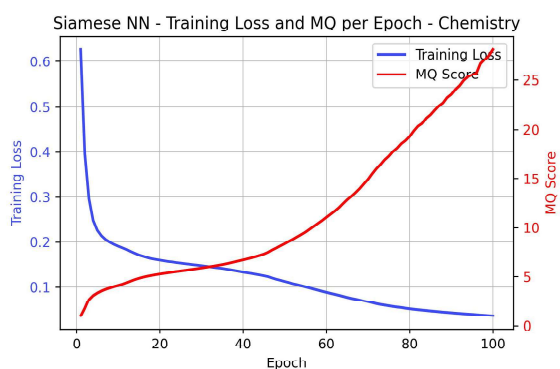


FIGURE 4.13: Training loss and MQ score for Siamese NN model - Chemistry

The training and test performance of the XGBoost model developed for AQI using Chemistry data are presented in the following figures. Figure 4.14 shows the training predictions and the prediction distribution and Figure 4.15 shows the test predictions and the prediction distribution.

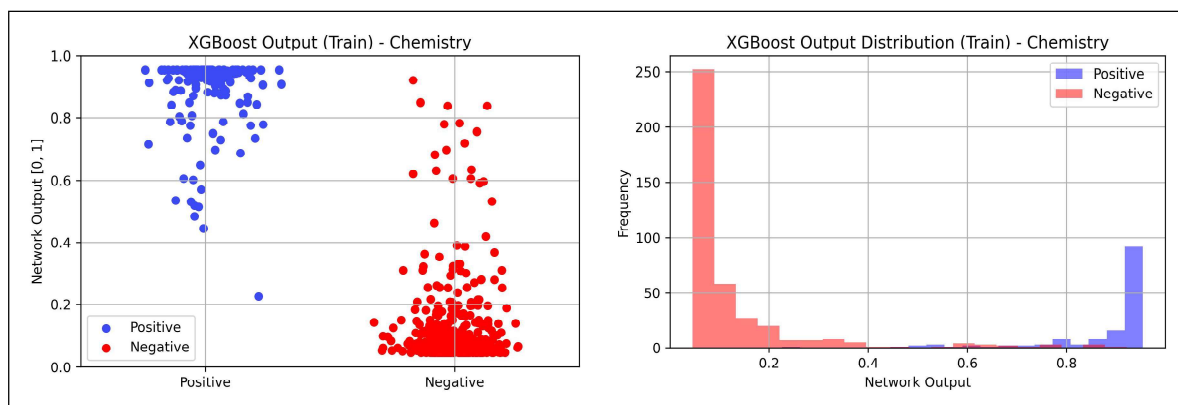


FIGURE 4.14: AQI distribution for XGBoost model (training) - Chemistry

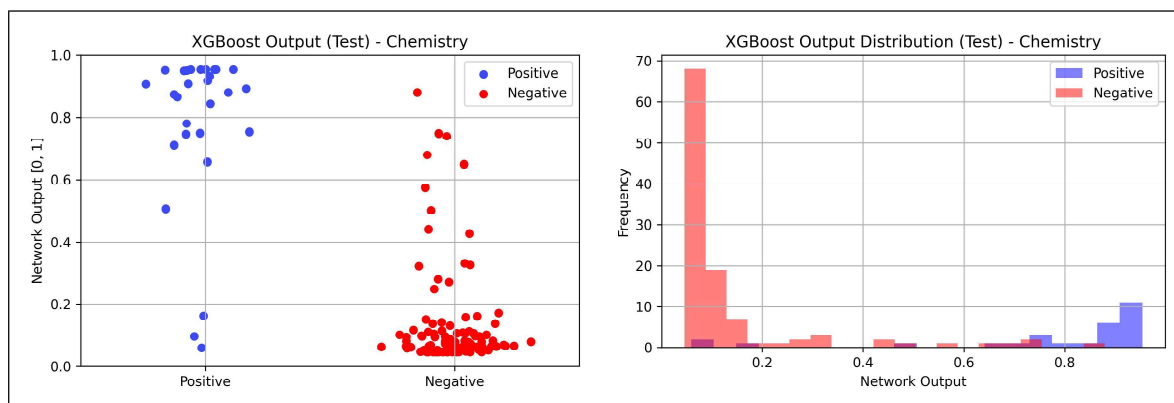


FIGURE 4.15: AQI distribution for XGBoost model (test) - Chemistry

4.2.3 Physiology or Medicine

The training and test performance of the Siamese NN model developed for AQI using Physiology or Medicine data is presented in the following figures. Figure 4.16 shows the training predictions and the prediction distribution, Figure 4.17 shows the test predictions and the prediction distribution, and Figure 4.18 displays the training loss and MQ score per epoch.

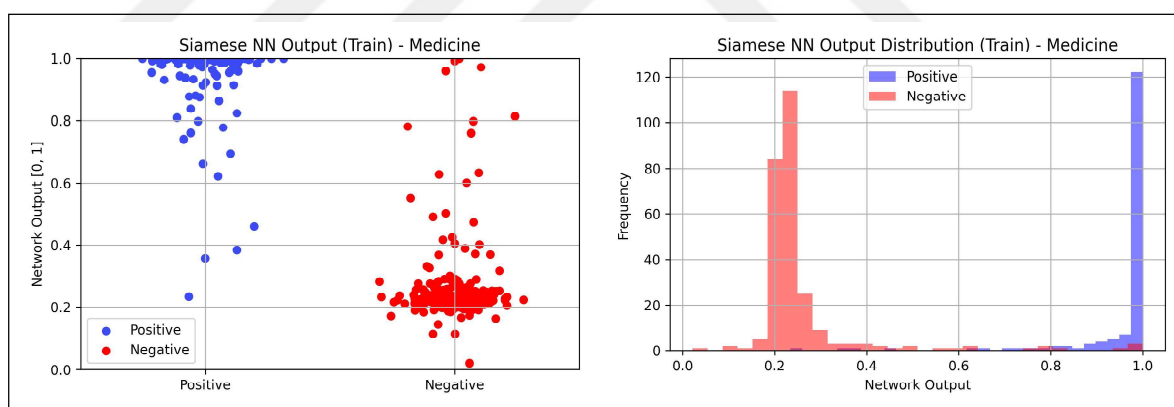


FIGURE 4.16: AQI distribution for Siamese NN model (training) - Physiology or Medicine

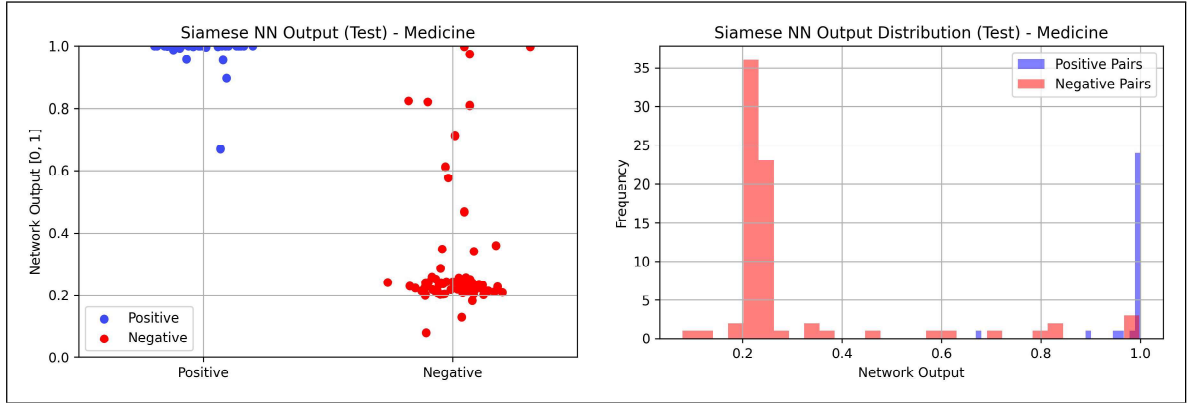


FIGURE 4.17: AQI distribution for Siamese NN model (test) - Physiology or Medicine

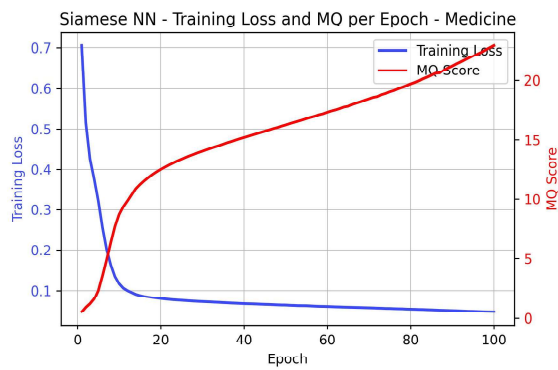


FIGURE 4.18: Training loss and MQ score for Siamese NN model - Medicine

The training and test performance of the XGBoost model developed for AQI using Physiology or Medicine data is presented in the following figures. Figure 4.19 shows the training predictions and the prediction distribution, and Figure 4.20 shows the test predictions and the prediction distribution.

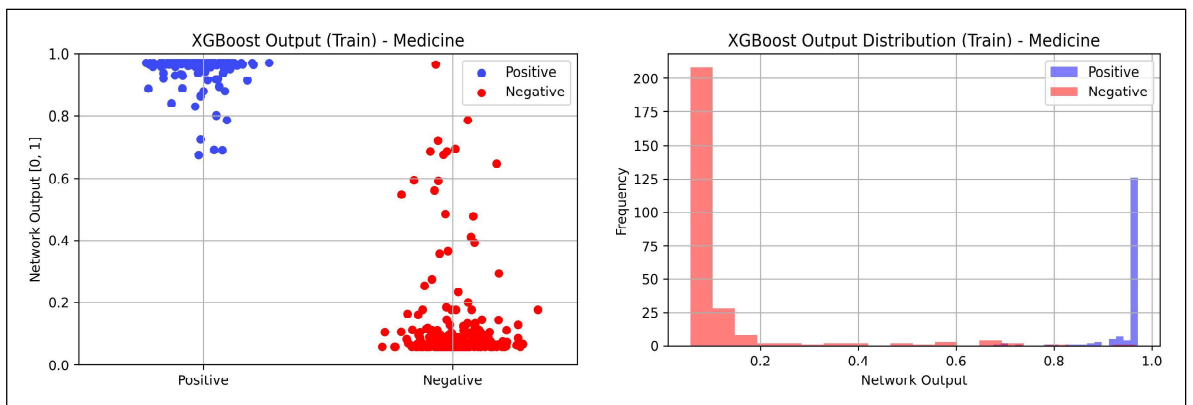


FIGURE 4.19: AQI distribution for XGBoost model (training) - Physiology or Medicine

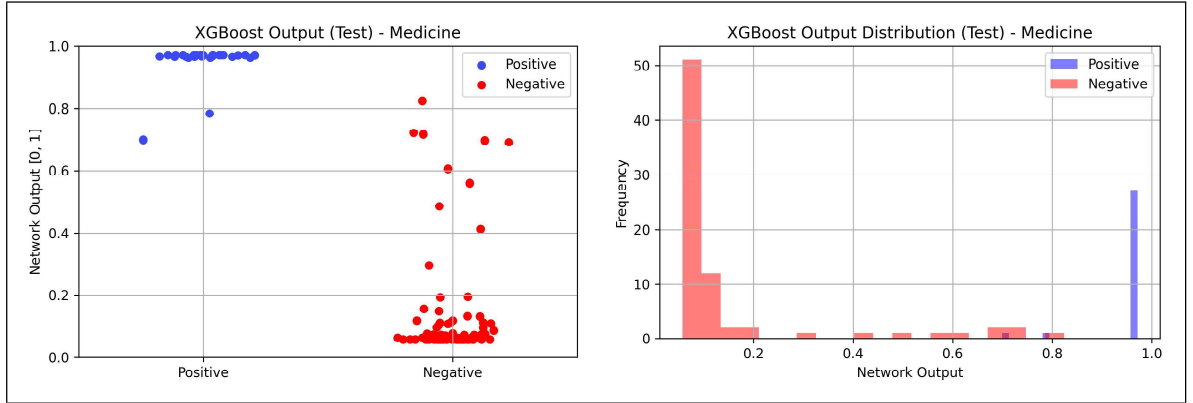


FIGURE 4.20: AQI distribution for XGBoost model (test) - Physiology or Medicine

4.2.4 Economics

The training and test performance of the Siamese NN model developed for AQI using Economics data is presented in the following figures. Figure 4.21 shows the training predictions and the prediction distribution, Figure 4.22 shows the test predictions and the prediction distribution, and Figure 4.23 displays the training loss and MQ score per epoch.

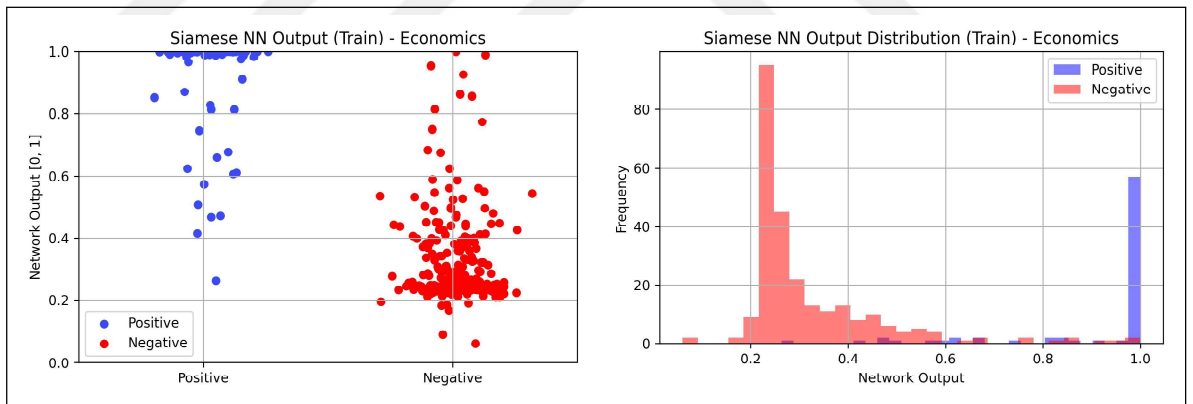


FIGURE 4.21: AQI distribution for Siamese NN model (training) - Economics

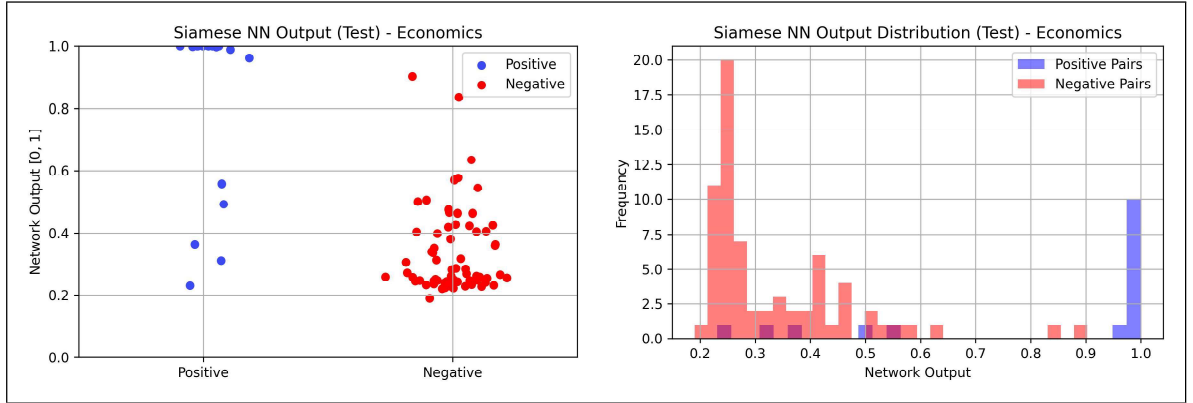


FIGURE 4.22: AQI distribution for Siamese NN model (test) - Economics

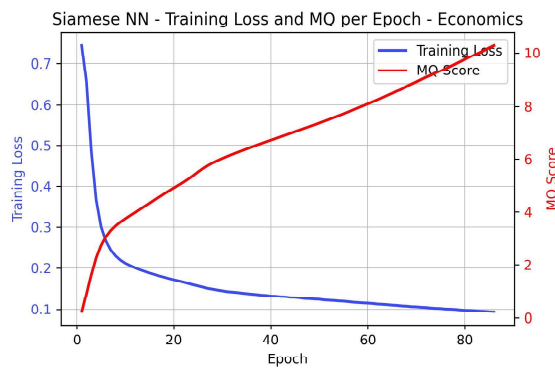


FIGURE 4.23: Training loss and MQ score for Siamese NN model - Economics

The training and test performance of the XGBoost model developed for AQI using Economics data is presented in the following figures. Figure 4.24 shows the training predictions and the prediction distribution, and 4.25 shows the test predictions and the prediction distribution.

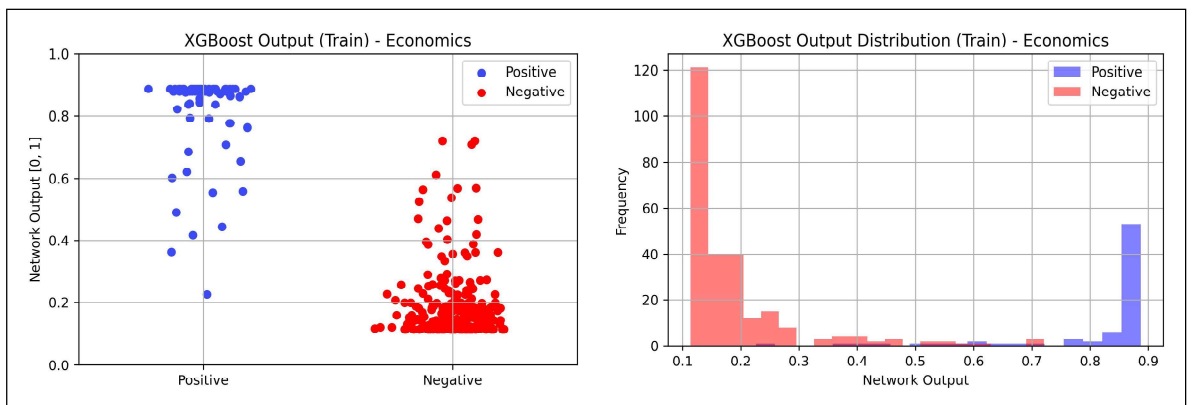


FIGURE 4.24: AQI distribution for XGBoost model (training) - Economics

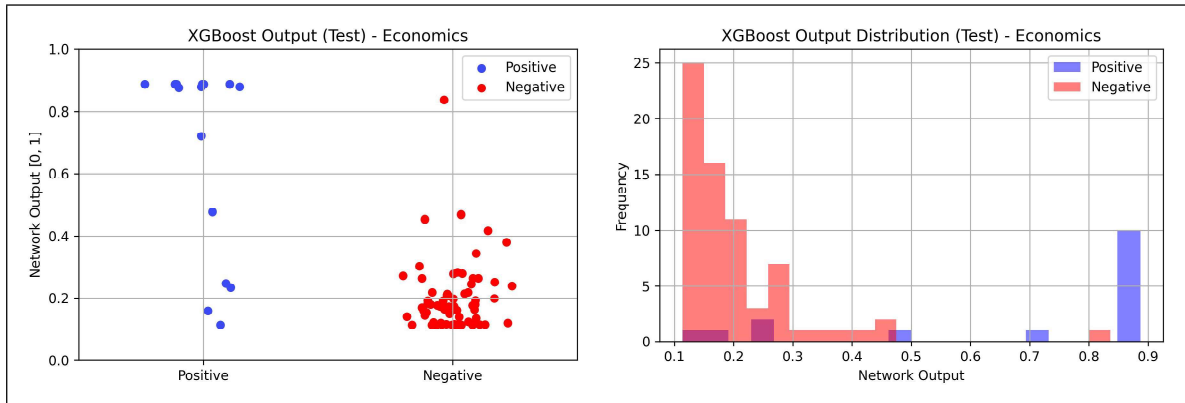


FIGURE 4.25: AQI distribution for XGBoost model (test) - Economics

4.3 Comparison of models

To evaluate the performance of the models, we use the MQ score. Tables 4.1, 4.2, 4.3, 4.4, and 4.5 present the MQ scores of all AQI prediction models.

TABLE 4.1: Comparison of training and test performance of models based on MQ score - All Fields.

Model type	Training/test	MQ
Siamese NN model	Training	13.8723
	Test	4.9099
XGBoost model	Training	9.0585
	Test	5.0824

TABLE 4.2: Comparison of training and test performance of models based on MQ score - Physics.

Model type	Training/test	MQ
Siamese NN model	Training	3.0464
	Test	2.1926
XGBoost model	Training	7.9752
	Test	3.2794

TABLE 4.3: Comparison of training and test performance of models based on MQ score - Chemistry.

Model type	Training/test	MQ
Siamese NN model	Training	5.9884
	Test	3.2639
XGBoost model	Training	15.4639
	Test	4.4942

Based on the results, the XGBoost models achieve the best performance, with the highest test MQ scores across both the full dataset and the field-specific subsets.

TABLE 4.4: Comparison of training and test performance of models based on MQ score - Physiology or Medicine.

Model type	Training/test	MQ
Siamese NN model	Training	15.8350
	Test	10.6731
XGBoost model	Training	31.6798
	Test	16.9994

TABLE 4.5: Comparison of training and test performance of models based on MQ score - Economics.

Model type	Training/test	MQ
Siamese NN model	Training	7.0383
	Test	2.1661
XGBoost model	Training	12.5909
	Test	2.1864

4.4 Feature Importances

Feature importance scores for the best performing models - XGBoost for both full dataset and field-specific subsets - are presented in Tables 4.6 and 4.7, and shown in Figures 4.26, 4.27 and 4.28. For the full dataset, feature importance scores for Siamese NN are also included in Figure 4.26 for comparison purposes.

TABLE 4.6: Feature importances for the best-performing model.

Feature importance scores					
Feature	Physics	Chemistry	Physiology or Medicine	Economics	All Fields
<i>cpppp-index</i>	24.40%	33.83%	50.12%	46.32%	31.32%
<i>a-index</i>	11.30%	3.11%	1.68%	4.45%	11.58%
<i>aap-index</i>	9.39%	20.75%	14.40%	5.50%	9.14%
<i>i100-index</i>	5.30%	5.20%	2.53%	0.00%	6.70%
<i>normalized h-index</i>	7.46%	5.57%	5.19%	4.64%	6.27%
<i>number of publications</i>	5.77%	3.89%	6.11%	5.09%	5.92%
<i>max_fa_citations</i>	5.21%	4.68%	3.43%	20.61%	5.37%
<i>cpp-index</i>	4.14%	4.53%	6.75%	3.98%	5.25%
<i>i1000-index</i>	3.58%	0.00%	0.00%	0.00%	4.81%
<i>h-index</i>	2.16%	3.24%	0.00%	1.61%	4.23%
<i>k-index</i>	3.28%	5.08%	6.23%	5.49%	2.94%
<i>i10-index</i>	6.32%	4.39%	3.00%	0.00%	2.71%
<i>Total citations</i>	6.96%	5.73%	0.00%	0.00%	2.16%
<i>g-index</i>	4.71%	0.00%	0.57%	2.31%	1.58%
<i>i10000-index</i>	0.00%	0.00%	0.00%	0.00%	0.00%

TABLE 4.7: Top 5 feature importance rankings.

Category	1st	2nd	3rd	4th	5th
All Fields	<i>cpppp-index</i>	<i>a-index</i>	<i>aap-index</i>	<i>i100-index</i>	<i>normalized h-index</i>
Physics	<i>cpppp-index</i>	<i>a-index</i>	<i>aap-index</i>	<i>normalized h-index</i>	<i>Total citations</i>
Chemistry	<i>cpppp-index</i>	<i>aap-index</i>	<i>Total citations</i>	<i>normalized h-index</i>	<i>i100-index</i>
Physiology or Medicine	<i>cpppp-index</i>	<i>aap-index</i>	<i>cpp-index</i>	<i>k-index</i>	<i>Number of publications</i>
Economics	<i>cpppp-index</i>	<i>max_fa_citations</i>	<i>aap-index</i>	<i>k-index</i>	<i>Number of publications</i>

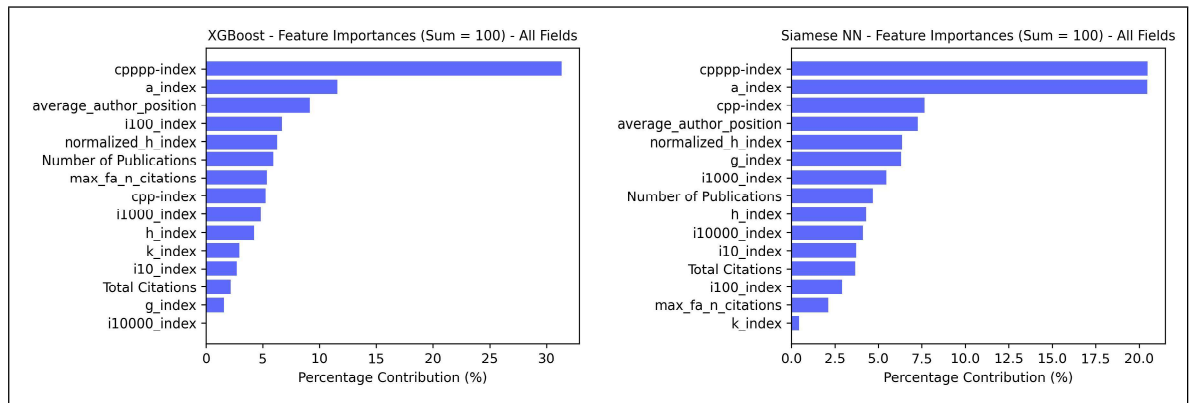


FIGURE 4.26: Feature importance scores for XGBoost and Siamese NN models - All Fields

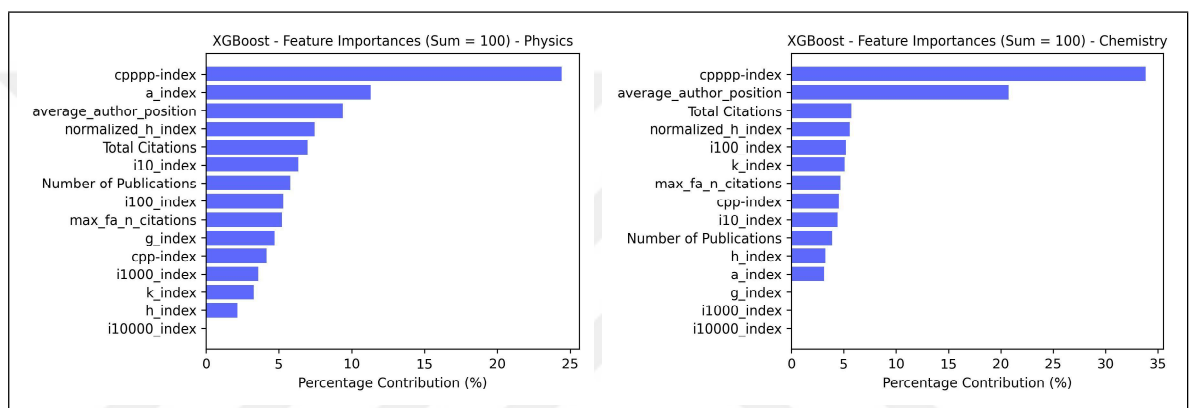


FIGURE 4.27: Feature importance scores for XGBoost model - Physics and Chemistry

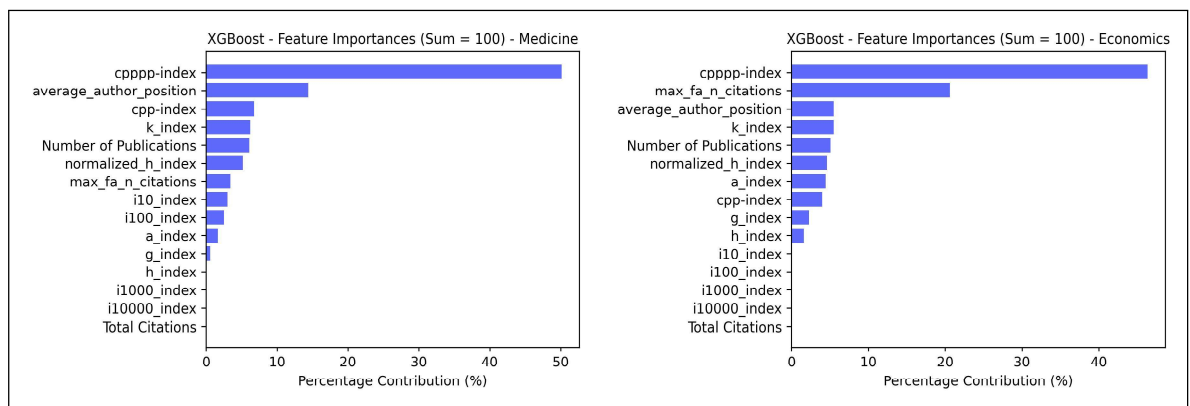


FIGURE 4.28: Feature importance scores for XGBoost model - Medicine and Economics

4.5 Discussion of results

4.5.1 Summary of findings

In line with the objective of this thesis, both models, the Siamese NN and XGBoost, were able to distinguish between positive and negative classes, predict AQI values, and

identify features that are strong indicators of academic quality both for the full dataset (all fields) and field-specific subsets. Table 4.8 presents the summary of results in the test data.

TABLE 4.8: Summary of test results for Siamese NN and XGBoost models.

Test	MQ Score	Accuracy* (Threshold = 0.5)	ROC AUC*	F1 Score*
Siamese NN:				
All Fields	4.9099	90.15%	94.59%	84.35%
Physics	2.1926	77.70%	89.19%	64.37%
Chemistry	3.2639	88.41%	95.84%	74.19%
Physiology or Medicine	10.6731	91.51%	99.01%	86.57%
Economics	2.1661	85.88%	88.77%	66.67%
XGBoost:				
All Fields	5.0824	92.08%	96.24%	86.45%
Physics	3.2794	92.81%	92.90%	84.38%
Chemistry	4.4942	92.75%	95.11%	83.33%
Physiology or Medicine	16.9994	93.40%	99.82%	89.23%
Economics	2.1864	92.94%	87.18%	78.57%

* The evaluation metric for the AQI is the MQ score, which measures the degree of separation between classes. Since AQI is a continuous value rather than a binary outcome, the ROC AUC, accuracy, and F1 score are reported for completeness and informational purposes but should be interpreted with caution. The evaluation focuses on the model's ability to quantify academic quality rather than to classify it into binary categories.

Models trained on the full dataset (all fields):

- Both models demonstrated a strong separation ability on the test data, with an average MQ of 5, ROC AUC > 0.9 , accuracy > 0.9 , indicating that chosen bibliometric indicators can meaningfully separate profiles of high academic quality from the average.
- XGBoost model outperformed Siamese NN in test data, with an MQ score of 5.08, ROC AUC of 0.96 and accuracy of 0.92.
- The top three feature importance ranking of the best performing model (XGBoost) are *cpppp-index*, *a-index* and *aap-index*. This supports findings from previous literature that citation intensity-based features are strong indicators of academic quality.

Models trained on field-specific subsets:

- All field-specific models demonstrated strong separation ability on the test data, indicating that chosen bibliometric indicators can meaningfully separate high academic quality profiles from the average.

- XGBoost models outperformed Siamese NN on the field-specific test datasets across all four fields, with an average MQ > 6 , ROC AUC > 0.90 , and accuracy > 0.90 .
- Field-specific models revealed differences in feature importance scores, highlighting variation in various disciplines. The top feature importance rankings of the best performing models include *cpppp-index*, *a-index*, *aap-index*, *max_fa_citations*, *Total citations*, *normalized h-index*, *k-index* and *i100-index*.

4.5.2 Interpretation

Model performance on AQI prediction:

- The results of both models demonstrate that bibliometric features used in this study can meaningfully distinguish between Nobel laureates and average researchers, with the positive class consistently receiving AQI scores close to 1.
- XGBoost: XGBoost, with its tree-based structure, provided the best performing models with strong test results. The ensemble learning models benefit from built-in techniques such as subsampling, tree-based regularization, and effective handling of class imbalance. These properties likely contribute to XGBoost's ability to work with small tabular datasets and to distinguish between the two classes more effectively. Additionally, XGBoost are less computationally expensive than Siamese NN and offer direct interpretability through built-in feature importance values.
- Siamese NN: The Siamese NN also successfully learned a meaningful classification of academic profiles, providing a robust complement to the XGBoost. The models are trained to maximize different-class separation while preserving same-class similarity using a modified contrastive loss function. However, their overall performance was slightly lower, likely due to overfitting, which is an inherent challenge when working with small datasets.
- Field-specific differences in model performance may indicate the distinct bibliometric characteristics of each discipline. For instance, in Physiology or Medicine, the model predictions are more accurate, suggesting that features used in this study more effectively capture academic quality in this field. Conversely, lower MQ scores observed in Physics and Economics may reflect limitations of citation-based

indicators in capturing academic quality in these fields, possibly due to varying publication behaviors.

Feature importance rankings:

Among the 15 features used in this study, feature importance scores are heavily concentrated among the top 3 to 5 features, whereas 1 to 5 features receive near-zero importance scores. The results reveal that there is a subset of bibliometric indicators, such as the *cpppp-index* and *a-index*, that consistently have strong predictive power for high academic quality.

All fields: The best performing model (XGBoost) highlights three features as strong indicators of academic quality, cumulatively accounting for more than 50% of the total feature importance:

- **cpppp-index**, a combined measure of both impact (citation) and individual contribution, received the highest weight with 31%, suggesting that the citation count and author count of the publications of a researcher are particularly important.
- **a-index**, a measure of average number of citations of the most influential work (h-core) of a researcher, is one of the most important features. While *h-index* is a combined measure of productivity and impact, it is not able to differentiate between researchers with the same h-index but varying levels of impact of h-core publications. This metric highlights the importance of consistent high-impact research over time and its top ranking aligns with its widespread recognition in scientometrics research.
- **aap-index**, a measure of individual contribution, is also one of the most important features capturing patterns in co-authorship roles. In many disciplines, author position interpreted as the level of involvement in the research, with first or low author position often signifies major contribution or leadership. The high importance score of this feature highlights the value of individual contribution in academic quality.
- Features such as **normalized h-index** and **i100-index** are also among the top-ranking contributors to the prediction of AQI. Although these are conventional

metrics of productivity and impact, their lower importance relative to contribution-sensitive features suggests that quantity alone is insufficient to capture academic quality.

- Notably, the top feature importance rankings of the XGBoost model are closely aligned with the Siamese NN model, strengthening confidence in the model's findings. Across both models, bibliometric features, *cpppp-index*, *a-index*, *aap-index* and *normalized h-index*, consistently emerged as the most important predictors of high academic quality.

Field-specific: In field-specific models, *cpppp-index* received the highest weight across all four fields in line with the model with full dataset (all fields). The rest of the feature importances were subtly changed, reflecting how different disciplines can value different aspects of the research output.

- **Physics:** Higher weight on **cpppp-index**, **a-index** and **aap-index** suggests that individual contribution, author position and impact of h-core publications are more critical in defining academic quality.
- **Chemistry:** Higher weight on **cpppp-index**, **aap-index** and **Total citations**, cumulatively accounting for 60% of the total feature importance, with more emphasis on author's position and contribution.
- **Physiology or Medicine:** Similar to Chemistry but shows more dominance in individual contribution, **cpppp-index** and **aap-index**, which cumulatively account for 65% of the total feature importance. **cpp-index** and **k-index** are also among the top ranking features, highlighting the importance of impact per publication.
- **Economics:** Higher weight on **cpppp-index**, **max_fa_citations** and **aap-index** suggests that impact and individual contribution are more critical.
- Features such as **k-index**, **i100-index**, **normalized h-index**, **number of publications** and **max_fa_citations**, showed more nuanced importance in field-specific analyses, illustrating that field-specific models are crucial for accuracy in academic evaluation.

4.5.3 Implications

This thesis combines bibliometric metrics and machine learning-based modeling, contributing to research on academic quality assessment to complement fair and effective evaluation. Our findings offer several important contributions:

- **Modeling academic quality:** The ability to model and predict academic quality using machine learning models based on bibliometric features with reasonable performance suggests opportunities for data-driven academic assessment, which could support more objective evaluations, particularly when complemented with peer review.
- **Feature transparency and interpretability:** The convergence between the Siamese NN and XGBoost models in the top feature importance rankings enhances the reliability of the models and offers actionable insights into the most important characteristics of academic success in general and across disciplines.
- **Field differentiation:** The study demonstrates that field-specific modeling could lead to more accurate assessments due to variation in test performances (MQ scores) and feature importance rankings between disciplines.

4.5.4 Limitations

Despite the strong performance of the models, this study has several limitations:

- **Benchmark of academic quality:** Reducing academic quality to a binary classification (Nobel laureate or not) ignores the other dynamics of academic performance. Researchers who are not Nobel laureates may still exhibit high quality that may not be captured in this study, limiting generalizability. Moreover, some high AQI values appear in the negative class, potentially representing outliers, such as professors with an unusually high number of publications. In addition, the negative class was chosen based on university rankings, which is a limited criterion.
- **Class imbalance and small data size:** Nobel laureates represent a small class, and the larger number of samples in the negative class may bias the models toward the majority class, potentially affecting prediction accuracy. Furthermore, although

our average group includes researchers data from 117 universities worldwide, the small size of field-specific data may limit the generalizability of the results.

- **Period sensitivity and temporal bias:** Nobel laureates span from the early 1900s to the present, while all negative class samples represent currently active researchers. Although selecting only full professors for the negative class was intended to create a more balanced comparison, temporal differences can still influence feature values. For example, recent publications by current professors may have accrued fewer citations simply due to the limited time since publication. As a result, younger or more recently active scholars may be penalized due to shorter publication histories.
- **Popularity effect among Nobel laureates:** Potential popularity effect may be observed among Nobel laureates, where winning the Nobel prize increases their visibility, attracts more collaborations, and leads to higher citation counts in the years following the award. This introduces a potential bias, as a portion of the metrics may result from the award itself rather than solely reflecting research productivity and impact, thereby potentially inflating the perceived academic quality of Nobel laureates relative to other researchers. The accuracy and magnitude of this assumed effect could be explored in future research. A possible approach would involve conducting a pre-award analysis of the data to quantify the Nobel prize's impact on the features. Nonetheless, it is important to note that Nobel laureates typically demonstrate outstanding academic performance and citation impact well before receiving the award, which may mitigate, though not entirely eliminate, this bias.
- **Omitted features:** The current model may exclude some important bibliometric features, such as journal quality of the publications. Incorporating additional features could further improve the performance and interpretability of the model, representing a promising direction for future research.
- **Data source limitations:** Both Web of Science and Scopus can contain inaccuracies or incomplete publications for certain researchers, potentially distorting bibliometric features. For example, one reason some Nobel laureates may have low AQI scores is that the data for earlier laureates may be incomplete or unavailable

online, especially for those awarded in the early 1900s, leading to an underestimation of their academic impact. In addition, not all Nobel laureates may have a large number of publications. Furthermore, given the high volume of publications (296,234) analyzed in this study, it was not feasible to manually verify each record, which may further contribute to the limitations of data quality.



Chapter 5

Conclusion and Future Directions

This thesis aims to develop an AI-based approach for quantifying the academic quality of researchers across various fields - Physics, Chemistry, Physiology or Medicine, Economics, and the combined (all fields) dataset. Specifically, it seeks to predict *Academic Quality Index* using supervised learning by distinguishing between 632 Nobel laureates (positive class) and 1,699 researchers affiliated with average-ranking universities (negative class).

To achieve this, we train machine learning models using two frameworks, Siamese NN and XGBoost, based on a set of 15 bibliometric features to predict AQI. The performance of the models is evaluated on test datasets using a custom metric, the MQ score. Finally, we extract feature importance scores to assess the contribution of each feature in the AQI predictions made by the trained models.

According to the results, XGBoost models demonstrated stronger performance on both the training and test datasets, indicating higher generalization capability.

Notably, XGBoost achieved its highest performance in the Physiology or Medicine, possibly due to the greater suitability of the model architecture and input features for distinguishing between classes and quantifying AQI in this field.

Siamese NN models, while more complex, shows potential in distinguishing classes with MQ scores close to those of XGBoost models.

Table 5.1 presents a summary comparison between the Siamese NN and XGBoost models developed in this thesis.

TABLE 5.1: Summary comparison between the Siamese NN and XGBoost.

	Siamese NN	XGBoost
Input data	15-feature tabular data	15-feature tabular data
Output	AQI prediction	AQI prediction
Training objective	supervised pairwise learning	supervised learning
Loss function	modified contrastive loss	logistic loss
Evaluation metric	MQ score	MQ score
MQ score (full test dataset)	4.9099	5.0824
Model performance	prone to overfitting on small data	robust generalization; handles class imbalance effectively
Computational cost	higher with 4.3 hours (due to large number of pair combinations)	lower with 0.3 hours
Feature importance determination	perturbation (manual)	built-in (gain-based)
Top ranking feature	cpppp-index	cpppp-index
Hyperparameter tuning	yes	yes
Regularization	yes	yes
Model complexity	more complex	less complex

The analysis of feature importance scores provided further insights into the key drivers of AQI predictions. Among the features, the *cpppp-index* consistently emerged as the most important across fields as well as the full dataset. Moreover, citation-sensitive features frequently emerged as strong indicators of academic quality.

Feature importance analysis also indicates nuanced differences in rankings across field-specific datasets, emphasizing the value of developing field-specific models to capture domain-specific characteristics.

All in all, the findings of this study offer data-driven insights that can support academic evaluations and decision making in academic recruitment. By complementing peer review with quantitative metrics, the approach presented in this thesis may enhance the fairness and consistency of researcher evaluations.

Future research can further explore academic quality quantification by using;

- alternative modeling approaches such as optimization- based and maximum-margin classifiers,
- additional features, particularly to capture publication quality, such as journal quartiles (Q1-Q4), reflecting the impact and prestige of journals,
- larger or more diverse datasets to enhance model generalizability,
- a pre-award analysis of Nobel laureates data to better understand and mitigate the bias introduced by receiving prestigious awards.

Appendix A

University List

A.1 List of universities for the average class researchers

The universities with which researchers in the average class are affiliated, which we consider to represent an average ranking, are listed in Table A.1. These universities are ranked between 601 and 800 in the 2025 Times Higher Education World University Rankings.

TABLE A.1: List of universities for the average class researchers.

#	University	Country
1	Al-Ahliyya Amman University	Jordan
2	Aligarh Muslim University	India
3	American University	United States
4	American University in Cairo	Egypt
5	American University of Sharjah	United Arab Emirates
6	Amity University, Noida	India
7	Auburn University	United States
8	Banaras Hindu University	India
9	Bar-Ilan University	Israel
10	Baylor University	United States
11	Beirut Arab University	Lebanon
12	Bilkent University	Turkey
13	Birla Institute of Technology and Science, Pilani	India
14	Boğaziçi University	Turkey

Continued on next page

Table A.1 – List of universities for the average class researchers (cont.).

#	University	Country
15	Chang Gung University	Taiwan
16	Chulalongkorn University	Thailand
17	Clark University	United States
18	COMSATS University Islamabad	Pakistan
19	Cyprus International University	Northern Cyprus
20	De Montfort University	United Kingdom
21	Eastern Mediterranean University	Northern Cyprus
22	Ewha Womans University	South Korea
23	Florida Atlantic University	United States
24	Fuzhou University	China
25	Gachon University	South Korea
26	Golestan University	Iran
27	Graz University of Technology	Austria
28	Hiroshima University	Japan
29	Imam Mohammad Ibn Saud Islamic University	Saudi Arabia
30	Indian Institute of Technology Patna	India
31	Iran University of Medical Sciences	Iran
32	Isfahan University of Technology	Iran
33	Jagiellonian University	Poland
34	Jiangnan University	China
35	Kafrelsheikh University	Egypt
36	Kansas State University	United States
37	Keio University	Japan
38	KIIT University	India
39	King Faisal University	Saudi Arabia
40	Kobe University	Japan
41	Lehigh University	United States
42	Liverpool John Moores University	United Kingdom
43	Louisiana State University	United States
44	Lovely Professional University	India
45	Majmaah University	Saudi Arabia
46	Malaviya National Institute of Technology	India
47	Manchester Metropolitan University	United Kingdom
48	Mansoura University	Egypt
49	Masaryk University	Czech Republic
50	Mazandaran University of Medical Sciences	Iran

Continued on next page

Table A.1 – List of universities for the average class researchers (cont.).

#	University	Country
51	Mississippi State University	United States
52	Nanjing Tech University	China
53	Nantes Université	France
54	National Sun Yat-Sen University	Taiwan
55	National University of Sciences and Technology	Pakistan
56	Near East University	Northern Cyprus
57	Nottingham Trent University	United Kingdom
58	Ohio University	United States
59	Ohio University (Main campus)	United States
60	Oklahoma State University	United States
61	Pontifical Catholic University of Rio de Janeiro	Brazil
62	Prince Sattam Bin Abdulaziz University	Saudi Arabia
63	Panjab University	India
64	Rochester Institute of Technology	United States
65	RUDN University	Russian Federation
66	Sultan Qaboos University	Oman
67	SUNY Binghamton University	United States
68	Taif University	Saudi Arabia
69	Tecnologico de Monterrey	Mexico
70	Texas Tech University	United States
71	The Catholic University of America	United States
72	The Catholic University of Korea	South Korea
73	The University of Jordan	Jordan
74	The University of Texas at San Antonio	United States
75	Toronto Metropolitan University	Canada
76	UiT The Arctic University of Norway	Norway
77	Universidade Federal do Rio Grande do Sul	Brazil
78	University of L'Aquila	Italy
79	University of Lincoln	United Kingdom
80	University of Malakand	Pakistan
81	University of Maryland, Baltimore County	United States
82	University of Minho	Portugal
83	University of Nevada, Las Vegas	United States
84	University of New Brunswick UNB	Canada
85	University of North Carolina at Charlotte	United States
86	University of Palermo	Italy

Continued on next page

Table A.1 – List of universities for the average class researchers (cont.).

#	University	Country
87	University of Parma	Italy
88	University of Pretoria	South Africa
89	University of Rennes 1	France
90	University of Sassari	Italy
91	University of Stavanger	Norway
92	University of Tabriz	Iran
93	University of Texas at Arlington	United States
94	University of the Basque Country	Spain
95	University of the Sunshine Coast	Australia
96	University of the Western Cape	South Africa
97	University of the West of Scotland	United Kingdom
98	University of Toledo	United States
99	University of Warsaw	Poland
100	University of Wyoming	United States
101	VIT University	India
102	Wenzhou University	China
103	William & Mary	United States
104	Worcester Polytechnic Institute	United States
105	Xi'an Jiaotong-Liverpool University	China
106	Xidian University	China
107	Yeungnam University	South Korea
108	Zhejiang University of Technology	China
109	Universiti Pendidikan Sultan Idris	Malaysia
110	Universiti Putra Malaysia	Malaysia
111	University of Alaska Fairbanks	United States
112	University of Aveiro	Portugal
113	University of Bari Aldo Moro	Italy
114	University of Denver	United States
115	University of Cagliari	Italy
116	University of Calabria	Italy
117	University of Crete	Greece

Appendix B

Pseudocode

This appendix contains the pseudocode for the Siamese NN and XGBoost models, providing a description of the algorithms.

Load and preprocess data:

- Assign binary labels: Nobel as 1 (positive class) and average as 0 (negative class)
- Split dataset into features (X) and labels (y)
- Split dataset into 80% training, 20% test

Note: Data split is performed once and used in all Siamese NN and XGBoost models (both for the models with full dataset and models with field-specific subsets).

B.1 Siamese neural network model

Note: Section below performed in all Siamese NN models (both for the model with full dataset and four models with field-specific subsets).

Standardize the features (mean=0, std=1)

Define build_pairs function:

- Function `build_pairs(positive_data, negative_data):`

- Takes two datasets: `positive_data` and `negative_data`
- Generate all possible pairs between each positive and each negative sample using a Cartesian product (via `np.meshgrid`)
- Return:
 - Array of repeated positive samples (`all_positive_pairs`)
 - Array of negative samples paired with them (`all_negative_pairs`)
 - The total number of such pairs

Example: Pair construction

Assume a 2-feature dataset with 2 positive class samples and 3 negative class samples as follows:

- **Positive class:**
 - Sample 1: [5, 6]
 - Sample 2: [7, 8]
- **Negative class:**
 - Sample 1: [1, 2]
 - Sample 2: [3, 4]
 - Sample 3: [2, 3]

Then, pairs using the `build_pairs` function would be formed as:

Pair #	Positive Sample	Negative Sample
1	[5, 6]	[1, 2]
2	[5, 6]	[3, 4]
3	[5, 6]	[2, 3]
4	[7, 8]	[1, 2]
5	[7, 8]	[3, 4]
6	[7, 8]	[2, 3]

Define a Siamese neural network:

```
Class SiameseNet(nn.Module):
```


- Input: `input_dim`
- Layers:
 - `fc1 = Linear(input_dim  $\rightarrow$  48) + ReLU + Dropout( $p = 0.3$ )`
 - `fc2 = Linear(48  $\rightarrow$  48) + ReLU`
 - `fc3 = Linear(48  $\rightarrow$  48) + ReLU`
 - `output = Linear(48  $\rightarrow$  1) + Sigmoid (binary probability)`
- Forward pass returns sigmoid output (probability for binary classification)

Define `contrastive_loss` function:

Define the function `contrastive_loss(model, positive_pairs, negative_pairs, margin_contrastive, margin_order, lambda_reg)`:

Inputs:

- `model`: the Siamese network (`SiameseNet`)
- `positive_pairs`: output of `build_pairs()` function, all training pairs composed of samples labeled as 1
- `negative_pairs`: output of `build_pairs()` function, all training pairs composed of samples labeled as 0
- `margin_contrastive` (m_c): float, margin for contrastive loss
- `margin_order` (m_o): float, margin for ordering constraint
- `lambda_reg` (λ): float, regularization coefficient

Forward pass both positive and negative pairs:

- `fP = model(positive_pairs)`
- `fN = model(negative_pairs)`
- Extract unique forward outputs as `fP_unique`, `fN_unique`

Compute loss components:

- Use `contrastive_loss` to minimize same-class variance:
- $\text{contrastive_loss} = \text{Var}(f_P_unique) + \text{Var}(f_N_unique) + \text{mean} \left(\max(0, m_c - |f_P - f_N|)^2 \right)$
- Use `order_loss` to ensure positive scores $>$ negative scores:
- $\text{order_loss} = \text{mean}(\max(0, m_o - (f_P - f_N)))$

`total_loss = contrastive_loss + order_loss`

Return `total_loss`

Define MQ function:

Define the function `compute_mq_score(model, positive outputs, negative outputs)`:

- Calculate MQ score for a given output
- $\text{MQ} = \frac{(\text{mean}(f_P) - \text{mean}(f_N))^2}{\text{var}(f_P) + \text{var}(f_N)}$

Perform grid search over margins (`margin_contrastive` and `margin_order`):

Define lists of contrastive and order margins:

- `margins_contrastive = [0.25, 0.5, 0.75]`
- `margins_order = [0.2, 0.3, 0.4, 0.5]`

Split training data again into training (75%) and validation (25%).

For each combination of `margin_contrastive` and `margin_order`:

- Generate positive/negative training and validation pairs
- Initialize SiameseNet and optimizer
 - `model = SiameseNet(input_dim)`
 - `optimizer = SGD(lr=1e-3)`
- Train model using mini-batch SGD:
- For epoch in range(`num_epochs`):

- Shuffle training pair indices
- For each batch:
 - Select batch_positive and batch_negative
 - Compute loss
 - Backpropagation and optimizer step
 - Compute train loss
- After final epoch:
 - Compute validation loss and MQ score
- Return MQ score and trained model

Select best margins as a result of grid search:

- Select best margins (`best_margin_contrastive`, `best_margin_order`) with highest validation MQ

Note: Section below performed separately for all Siamese NN models (one model with full dataset and four models with field-specific subsets).

Perform early stopping for best number of epochs: Model training with k-fold cross-validation (k=5).

- `maximum_number_epochs = 100`
- `patience = 10 epochs`

Perform 5-fold CV using best margins on the full training set to calculate the best number of epochs: For each fold:

- Split train/validation sets
- Build positive/negative pairs
- Initialize model and optimizer

- For each epoch (up to `maximum_number_epochs`):
 - Train on training pairs
 - Evaluate on validation pairs
 - Track `best_validation_MQ`
 - If `best_validation_MQ` improves: save `state_dict`
 - Else: increase `no_improvement` counter
 - If `no_improvement` $\geq$ `patience`: stop early and save `num_epochs`
- Save loss and MQ history per fold

After all folds are completed:

- Select fold with highest average validation MQ score
- Set `num_epochs` as the `best_number_epochs`

Final model training with best margins and best number of epochs:

- Create all possible positive-negative pairs from full training set
- Train final model for `best_number_epochs` and best margins (`best_margin_contrastive`, `best_margin_order`)
- Track training loss and MQ score at each epoch

Feature importance computation: Calculate feature importance scores via zero-out perturbation:

- Get baseline model output on original data: `base_outputs = model(X_train)`
- For each feature `i`:
 - Set column `i` to 0 (perturb) across all samples
 - Measure updated output: `perturbed_outputs`
 - Compute mean absolute change in output as:
 - `feature_importance = mean(abs(base_outputs - perturbed_outputs))`

- Normalize importance scores to sum to 1

Evaluate model output (AQI predictions) for training and test sets:

Training data:

- `AQI_positive = final_model(positive_data_train)`
- `AQI_negative = final_model(negative_data_train)`
- Calculate MQ score

Test data:

- `test_AQI_positive = final_model(positive_data_test)`
- `test_AQI_negative = final_model(negative_data_test)`
- Calculate MQ score

B.2 XGBoost model

Note: Section below performed separately for all XGboost models (one model with full dataset and four models with field-specific subsets):

Prepare data: Use the same train/test dataset used in Siamese NN model

Model tuning via grid search:

Define grid search parameters:

- `n_estimators: [30, 35, 40, 45, 50]` *Lower number of trees for small dataset*
- `max_depth: [3, 5, 7]` *Shallow trees for generalization*
- `learning_rate: [0.01, 0.05]` *Small learning rates*
- `subsample: [0.8, 1]` *Random row sampling*
- `min_child_weight: [3, 5, 7, 10]` *Node split regularization*

- `reg_lambda: [1, 5, 10]` *L2 regularization (smooth weights)*
- `reg_alpha: [0.01, 0.1]` *L1 regularization (sparse models)*
- `scale_pos_weight: [1, 2, 2.7, 2.8, 3, 3.6, 4]` *Class balancing*

Define base model:

- `XGBClassifier(eval_metric="logloss", random_state=42)`

Define cross-validation:

- `cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)`

Define grid search parameters :

```
grid_search = GridSearchCV(  
    estimator=xgb_base,  
    param_grid=param_grid,  
    scoring=mq_score_func_cv,  
    cv=cv,  
    verbose=1,  
    n_jobs=1)
```

Perform grid search:

```
grid_search.fit(X_train, y_train)
```

Output:

- Best hyperparameters: `grid_search.best_params_`
- Best cross-validated MQ score: `grid_search.best_score_`

Evaluate model output (AQI predictions) for training and test sets:

Retrieve the final model:

- `best_xgb = grid_search.best_estimator_`

Training set:

- `y_train_proba = best_xgb.predict_proba(X_train)[:, 1]`
- `AQI_positive = y_train_proba[y_train == 1]`
- `AQI_negaiive = y_train_proba[y_train == 0]`
- Calculate MQ score

Test set:

- `y_test_proba = best_xgb.predict_proba(X_test)[:, 1]`
- `test_AQI_positive = y_test_proba[y_test == 1]`
- `test_AQI_negative = y_test_proba[y_test == 0]`
- Calculate MQ score

Feature importance computation:

- `feature importances = best_xgb.feature_importances_`

Bibliography

- [1] Ajay Badugu, KS Arunab, and Aneesh Mathew. Predicting land surface temperature using data-driven approaches for urban heat island studies: a comparative analysis of correlation with environmental parameters. *Modeling Earth Systems and Environment*, 10(1):1043–1076, 2024.
- [2] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a " siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.
- [3] Times Higher Education. World university rankings 2025, 2025. <https://www.timeshighereducation.com/world-university-rankings/latest/world-ranking> (Accessed: 2025-06-20).
- [4] Pandi Chellappandi and Channaveerapla S Vijayakumar. Bibliometrics, scientometrics, webometrics/cybermetrics, informetrics and altmetrics—an emerging field in library and information science research. *Shanlax International Journal of Education*, 7(1):5–8, 2018.
- [5] Lutz Bornmann and Loet Leydesdorff. Scientometrics in a changing research landscape: bibliometrics has become an integral part of research quality evaluation and has been changing the practice of research. *EMBO reports*, 15(12):1228–1232, 2014.
- [6] Melissa K McBurney and Pamela L Novak. What is bibliometrics and why should you care? In *Proceedings of the International Professional Communication Conference*, pages 108–114, Portland, OR, USA, 2002.
- [7] Philip M Podsakoff, Scott B MacKenzie, Nathan P Podsakoff, and Daniel G Bachrach. Scholarly influence in the field of management: A bibliometric analysis

- of the determinants of university and author impact in the management literature in the past quarter century. *Journal of management*, 34(4):641–720, 2008.
- [8] Jorge E Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.
- [9] Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.
- [10] Bihui Jin. H-index: an evaluation indicator proposed by scientist. *Science Focus*, 1(1):8–9, 2006.
- [11] Bihui Jin, LiMing Liang, Ronald Rousseau, and Leo Egghe. The r-and ar-indices: Complementing the h-index. *Chinese science bulletin*, 52(6):855–863, 2007.
- [12] Qiang Wu. The w-index: A significant improvement of the h-index. *arXiv preprint arXiv:0805.4650*, 2008.
- [13] Sergio Alonso, Francisco Cabrerizo, Enrique Herrera-Viedma, and Francisco Herrera. hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. *Scientometrics*, 82(2):391–400, 2010.
- [14] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and technology*, 59(5):830–837, 2008.
- [15] Chun-Ting Zhang. The e-index, complementing the h-index for excess citations. *PloS one*, 4(5):e5429, 2009.
- [16] Pablo D Batista, Mônica G Campiteli, and Osame Kinouchi. Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1):179–189, 2006.
- [17] Lorna Wildgaard, Jesper W Schneider, and Birger Larsen. A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, 101:125–158, 2014.
- [18] Muhammad Usman, Ghulam Mustafa, and Muhammad Tanvir Afzal. Ranking of author assessment parameters using logistic regression. *Scientometrics*, 126(1):335–353, 2021.

- [19] Anand Bihari and Sudhakar Tripathi. Em-index: a new measure to evaluate the scientific impact of scientists. *Scientometrics*, 112:659–677, 2017.
- [20] Anthony FJ Van Raan. Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67:491–502, 2006.
- [21] Thed Van Leeuwen, Martijn Visser, Henk Moed, Ton Nederhof, and Anthony Van Raan. The holy grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, 57(2):257–280, 2003.
- [22] Ben Martin. The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3):343–362, 1996.
- [23] Giuliano Franco. Scientific research of senior italian academics of occupational medicine: a citation analysis of products published during the decade 2001–2010. *Archives of Environmental & Occupational Health*, 70(2):110–115, 2015.
- [24] John PA Ioannidis, Richard Klavans, and Kevin W Boyack. Multiple citation indicators and their composite across scientific disciplines. *PLoS biology*, 14(7):e1002501, 2016.
- [25] Massimo Franceschet. A cluster analysis of scholar and journal bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 60(10):1950–1964, 2009.
- [26] Lorna Wildgaard. A critical cluster analysis of 44 indicators of author-level performance. *Journal of Informetrics*, 10(4):1055–1078, 2016.
- [27] Andrea Giovanni Nuzzolese, Paolo Ciancarini, Aldo Gangemi, Silvio Peroni, Francesco Poggi, and Valentina Presutti. Do altmetrics work for assessing research quality? *Scientometrics*, 118(2):539–562, 2019.
- [28] Lutz Bornmann and Robin Haunschild. Do altmetrics correlate with the quality of papers? a large-scale empirical study based on f1000prime data. *PloS one*, 13(5):e0197133, 2018.
- [29] Pablo Jensen, Jean-Baptiste Rouquier, and Yves Croissant. Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, 78(3):467–479, 2009.

- [30] Jason R Tregellas, Jason Smucny, Donald C Rojas, and Kristina T Legget. Predicting academic career outcomes by predoctoral publication record. *PeerJ*, 6:e5707, 2018.
- [31] Bilal Ahmed, Li Wang, Ahmad Sami Al-Shamayleh, Muhammad Tanvir Afzal, Ghulam Mustafa, Wagdi Alrawagfeh, and Adnan Akhunzada. Machine learning approach for effective ranking of researcher assessment parameters. *IEEE Access*, 11:133294–133312, 2023.
- [32] Francesco Poggi, Paolo Ciancarini, Aldo Gangemi, Andrea Giovanni Nuzzolese, Silvio Peroni, and Valentina Presutti. Predicting the results of evaluation procedures of academics. *PeerJ Computer Science*, 5:e199, 2019.
- [33] Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos. Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280, 2007.
- [34] Fred Ye and Ronald Rousseau. Probing the h-core: an investigation of the tail–core ratio for rank distributions. *Scientometrics*, 84(2):431–439, 2010.
- [35] Jaime A Teixeira da Silva. The i100-index, i1000-index and i10, 000-index: Expansion and fortification of the google scholar h-index for finer-scale citation descriptions and researcher classification. *Scientometrics*, 126(4):3667–3672, 2021.
- [36] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, San Francisco, CA, USA, 2016.
- [37] Davide Chicco. Siamese neural networks: An overview. *Artificial neural networks*, pages 73–94, 2021.
- [38] Yikai Li, CL Philip Chen, and Tong Zhang. A survey on siamese network: Methodologies, applications, and opportunities. *IEEE Transactions on artificial intelligence*, 3(6):994–1014, 2022.
- [39] Hiteshi Jain, Gaurav Harit, and Avinash Sharma. Action quality assessment using siamese network-based deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2260–2273, 2020.

- [40] Michal Byra, Katarzyna Dobruch-Sobczak, Ziemowit Klimonda, Hanna Piotrkowska-Wroblewska, and Jerzy Litniewski. Early prediction of response to neoadjuvant chemotherapy in breast cancer sonography using siamese convolutional neural networks. *IEEE journal of biomedical and health informatics*, 25(3):797–805, 2020.
- [41] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany, 2016.
- [42] Swati Jindal, Gaurav Gupta, Mohit Yadav, Monika Sharma, and Lovekesh Vig. Siamese networks for chromosome classification. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 72–81, Venice, Italy, 2017.
- [43] Qi Wang, Junyu Gao, and Yuan Yuan. Embedding structured contour and location prior in siamesed fully convolutional networks for road detection. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):230–241, 2017.
- [44] Bin Wang and Dian Wang. Plant leaves classification: A few-shot learning method based on siamese network. *IEEE Access*, 7:151754–151763, 2019.
- [45] Linchang Zhao, Zhaowei Shang, Ling Zhao, Anyong Qin, and Yuan Yan Tang. Siamese dense neural network for software defect prediction with small data. *IEEE Access*, 7:7663–7677, 2018.
- [46] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2598–2607, Snowmass Village, CO, USA, 2020.
- [47] Neil Zeghidour, Gabriel Synnaeve, Nicolas Usunier, and Emmanuel Dupoux. Joint learning of speaker and phonetic similarities with siamese networks. In *INTER-SPEECH*, pages 1295–1299, San Francisco, CA, USA, 2016.
- [48] Kelwin Fernandes, Davide Chicco, Jaime S Cardoso, and Jessica Fernandes. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Computer Science*, 4:e154, 2018.

- [49] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30, Lille, France, 2015. Lille.
- [50] Md Johir Raihan, Md Al-Masrur Khan, Seong-Hoon Kee, and Abdullah-Al Nahid. Detection of the chronic kidney disease using xgboost classifier and explaining the influence of the attributes on the model using shap. *Scientific Reports*, 13(1):6263, 2023.
- [51] Sami Ben Jabeur, Nicolae Stef, and Pedro Carmona. Bankruptcy prediction using the xgboost algorithm and variable importance feature engineering. *Computational Economics*, 61(2):715–741, 2023.
- [52] Yung-Chia Chang, Kuei-Hu Chang, and Guan-Jhih Wu. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, 73:914–920, 2018.
- [53] Xiupeng Shi, Yiik Diew Wong, Michael Zhi-Feng Li, Chandrasekar Palanisamy, and Chen Chai. A feature learning approach based on xgboost for driving assessment and risk prediction. *Accident Analysis & Prevention*, 129:170–179, 2019.
- [54] Zhixun Zhao, Hui Peng, Chaowang Lan, Yi Zheng, Liang Fang, and Jinyan Li. Imbalance learning for the prediction of n 6-methylation sites in mrnas. *BMC genomics*, 19:1–10, 2018.