Babylonian Journal of Machine Learning
Vol.2025, **pp**. 1–12
DOI: https://doi.org/10.58496/BJML/2025/001; ISSN: 3006–5429
https://mesopotamian.press/journals/index.php/BJML

Research Article

Integrating Machine Learning and Genetic Algorithms to Enhance Gene-Disease Classification: An XBNet-Based Framework

Rana Khalid Hamad[1, *,]

[1] *Islamic university of Lebanon*

ARTICLEINFO

Article History
Received 15 Oct 2024
Revised: 14 Nov 2024
Accepted 15 Dec 2024
Published 10 Jan 2025

Keywords
Gene-Disease Classification
Machine Learning
Deep Learning
Genetic Algorithms
XBNet
Bioinformatics
Gene Expression

ABSTRACT

In bioinformatics, the classification of gene-disease associations is crucial. It directly affects whether we can untangle the genetic roots of various disease as well as if we will find some justifiable therapy for these cured diseases. Using XBNet to construct genetic algorithms for higher accuracy and speeds of gene-disease classification--this is the method developed in the book. Consisting of gene expression profiles for six diseases--Alzheimer's, Asthma, Cancer, Diabetes, Fabry and Down syndrome--our research has applied a comprehensive pre-processing technique to this data set from Kaggle. This has included such things as eliminating stop-words and punctuation marks and tokenization. Using the terms of Frequency (TF) and of Term Frequency-Inverse Document Frequency (TF-IDF method) for features extraction, our text data on genes are transformed into numerical axes fit for input to machine learning models.

1. INTRODUCTION

We used the XBNet model boosted gradient descent algorithm to train and optimize it using the importance of parameters In the current study specifications in order to make predictions. Meanwhile, two versions (V0 and V1) of the dataset were tried out and feature selections were carried out with both TF and TF-IDF methods. XBNet-related training and testing data on TF based selection led to the highest accuracy so far since XBNet came into being, 99% in training and 98% over its counterpart not yet set up for general consumption. But introduced to the same problem, TF-IDF-based selections showed disappointing results in some ways downright unpalatable. The classification accuracy indices of our results for TF and TF-IDF methods are further demonstration that feature selection and competitive learning can throw a spotlight on a common problem, calling into question how effective the current approach with XBNet mtx. Heretofore, integrating genetic algorithms with the method of feature selection not only effectively eliminates redundant information but also helps improvements in rationality and model generalization ability.Its result not only opens new methodologies in gene-disease classification methods but also offers us new ways to apply computation to data analysis within the context of bioinformatics, clearing the road for further studies on genetics with greater accuracy and reliability.The advent of high-throughput technologies in sequencing has transformed the field of bioinformatics: it can now generate huge amounts of gene microarray data all at once.

As pivotal tools in bioinformatics, Machine Learning (ML) and Deep Learning(DL) have brought with them robust ways of analyzing the most complex biological data [1-4]. In pattern-recognition, feature extraction,and predictive modeling, these techniques excel. Therefore they are essential for unraveling complexities nested deep within genomic datasets, such as those described in references 1. But the effectiveness of these models always depends upon data preprocessing quality and feature selection. The same holds true for model parameters optimization.

Consequently it is necessary to combine advanced ML and DL techniques with domain-specific algorithms in order to enhance the accuracy and reliability of models for classifying genes and diseases. And biology has been becoming straight

*Corresponding author. Email: it@gmail.com

and straight.On top of several crucial advances, a whole set of problems persists within gene-disease classification. For example, the high dimensionality of genetic data; the redundancy of feature states; or even different Bureau of Biological Processes make researchers struggle to build models 2! Further, which feature extraction and selection method to choose is also crucial to the performance of classification algorithms. For instance, while Term Frequency (TF) has been shown to be effective at capturing relevant gene expressions in certain situations, Term Frequency-Inverse Document Frequency (TF-IDF) may not yield as good results if used in another biological context 5.This paper presents a novel method that combines the XBNet architecture with genetic algorithms to optimize the selection of features and improve the performance of the model. Employing a substantial dataset from Kaggle, the study majors in genesof six different diseases: Alzheimer's, Asthma, Cancer, Diabetes, Fabry and Down Syndrome. Through meticulous data processing, feature extractionand the application of advanced ML techniques, it aims to create a sound framework for accurate gene-disease classification. The remainder of this dissertation will give a broader survey concerning currently available literature, expound the methodological framework engaged on and provide resultant data, plus discuss its implications for the wider field of bioinformatics and personalized medicine.

2- LITERATURE REVIEW

Various studies have played with the intersection of machine learning and bioinformatics to more clearly make any gene-disease link, using many different algorithm-selection techniques. These studies employ a variety of algorithms, techniques for feature selection, and strategies to optimize genetic data.

Dara et al. [1] author a well-read paper that reviews machine learning applications in drug design, and points out how ML techniques are profoundly changing how we can extrapolate new drugs. Their study shows how ML algorithms show the same flexibility as some lawnmowers and hay balers together with a bonus for any programmers: they fit well to handling different biological data sets. Different stages in drug synthesis demand distinct programming methods.

Piccolo et al. [2] explore variance in patient classification based on gene-expression data, showing that adoption of algorithm and score indicator will markedly affect the result. Their findings suggest that careful choice of machine learning models and some appropriate criteria for evaluation can guarantee both exactness and reproducibiity of classification results in clinical settings.

Gokulakrishnan et al. (3) apply deep learning techniques for predicting diseases on the basis of microarray data. Their work demonstrates the superiority of deep neural networks when it comes to complex gene expression patterns and therefore also how well these can predict various diseases. This work shows just how deep learning models should not to be underestimated even at the cost of traditional approaches because if they can handle the flighty, flickery character of high-dimensional genetic data.

Greener et al. [4] present a roadmap of sorts for biologists to install machine learning and discuss those fundamental concepts related to ML which biologists should know. This comprehensive overview provides useful guidance for any research workers intending to introduce ML technique into the study of any biological subject, and highlights the significance of interdisciplinary cooperation.

Text mining is an important field where bioinformatics intersects. The literature often has comprehensive reviews of this subject. The diversity of industries, with its various processes and data formats, makes it important for standard text mining techniques to be maintained across them. Skarpathiotaki, Psannis and Charalabidis [5] The consistency of methodologies in extracting meaningful insights from textual data is absolutely important, while Egger, and Gokce [6] make readers familiar with the essential knowledge of Coincident generation Natural Language Processing (NLP) techniques, which are basic for understanding and putting bioinformatics text mining applications into practice.

After a short survey by Ramanathan and Meyyappan [7] and Vidya, Aghila [8], Sagayam et al. [9], Gupta Lehal [10] lists Hearst [11], Agrawal Batra [12] Patel, Soni [13], Patel Sharma [14], the varied techniques and tools for text mining are emphasized collectively. They look at the differences between them in a new way such as retrieval and extraction, search engine indexing. Application of text mining methods has made it possible to analyse texts in several domains including healthcare bioinformatics. In this way their comprehensive analysis has illuminated the methodological advancements and problems in computer text processing,as well as be fitting for handling large-scale bioinformatics data sets. There's every reason to try and carry out intelligent image enhancement: According to Zortea Plaza [15] geographic pre-processing technology used for end member extraction, while Supriya Subaji [16] compares direct- and indirect-contrast enhancement techniques in figure photograph processing carefully.

Bioinformatics has its biological data represented many different ways. These studies relate to it is a proposition of image data much as the use histological slides and general cropping practices for trees or fractal pixels do not but administrations of life choice or accident turn into workable outputs from higher plants. One of the critical resources is large-scale gene nomenclators. He gives the full details in Braschi et al [17]. Homologous genes in different studies can be classified and identified consistently only with standardized gene nomenclature, which is why genenames.org Sarsdel et al. Snyder et al.

[18] work with large nucleotide sequences of soapberry Bug Virus (SBVV) and elucidation problematics in handle serological groups III resistance of GVA ligand resistance using a bacterial system for selection and gene cloning. makes all characteristics of gene names available online absolutely free of charge! Another example is provided by Snyder, who examines the genetic basis for clinical responses to immunotherapy in melanoma by using genetic data.

What can we learn from it?In recent studies, gene-disease classification has been given priority and ML, DL models' effectiveness is shown. Alrefaai Gatmir and Musa Alrashid. Two algorithms proposed for the classification of gene expression data The modified K-Nearest Neighbors (KNN) combined with Arti%cial Bee Colony (ABC) algorithmis used. For Type 1 Diabetes, the accuracy of the results is high. Alzoubi Assad et al. Multi-omics genetic variation deep learning framework for complex diseases risk prediction As a result, the prediction accuracy is significantly more accurate than traditional models in use. Jo Dong-Jae et al. Convolutional Neural Networks (CNN) combined with Bayesian optimization to classify Alzheimer's disease based on gene expression data The classification accuracy has been greatly enhanced. Qumsiyeh Radwan et al. A knowledge-based approach using an ML technique for uncovering gene associations across various diseases GediNET The increased precision is considerable. Sethi Jatin et al. By using Bayesian parameter optimization to optimize the deep learning model for Alzheimer's disease identification, accuracy rates were extremely high. Wu Yi and Yang Hicks In classifying of breast cancer type with genomic data, the effectiveness of the XGBoost algorithm model becomes evident. Ariani Anna et al The combination of Artificial Bee Colony and genetic modified KNN algorithms were used to classify kidney disease, with performance well surpassed any other method used. Gurovich Yaron et al. Tips on how to use deep learning with genetic disorders that can be seen on people’s faces The key message is that deep learning has a role to play in phenotype-genotype correlations. Asif Mahmud et al Using gene Ontology methodological maps to provide gene functional similarity of disease related genes, by means of machine learning integrated with the method The classification accuracy is greatly increased. All these studies taken together help to show that advanced machine learning technologies are being used in gene-disease classification in an increasingly rewarding way. Optimization algorithms, feature selection and various architectures of deep learning have all accelerated predictive accuracy and trustfulness in models tremendously, leading the way for better personalized medicine and more efficient therapies on a case-by-case basis.

3- METHODOLOGY

This paper delineates the methodological framework employed in this study to classify gene-disease associations using the XBNet architecture integrated with genetic algorithms. The methodology encompasses data acquisition, preprocessing, feature extraction, feature selection, model training, and evaluation.

3.1 Data Acquisition

The dataset utilized in this research was sourced from Kaggle, comprising gene expression profiles associated with six distinct diseases: Alzheimer’s, Asthma, Cancer, Diabetes, Fabry, and Down Syndrome. The dataset is provided in a CSV format, containing various parameters pertinent to each gene-disease association. The selection of this dataset is informed by its comprehensive coverage of multiple diseases and the availability of high-dimensional gene expression data, making it suitable for evaluating the proposed classification framework.

3.2 Data Preprocessing

Raw genetic data is inherently unstructured and may contain inconsistencies, missing values, and noise. Effective preprocessing is essential to ensure data quality and to enhance the performance of machine learning models [19][20]. The preprocessing steps undertaken in this study include:

- Data Cleaning: Removal of missing values, duplicates, and irrelevant features to streamline the dataset.
- Normalization: Scaling gene expression values to a standard range to ensure uniformity across different features.
- Feature Transformation: Applying transformations such as log-scaling to manage skewed distributions and to stabilize variance [21][22].

3.3 Feature Extraction

Feature extraction is a critical step in transforming raw gene expression data into a format suitable for machine learning models. Two primary feature extraction methods were employed:

- Term Frequency (TF): Calculates the frequency of each gene expression within the dataset, providing a straightforward representation of gene activity [23][24].
- Term Frequency-Inverse Document Frequency (TF-IDF): Weighs the frequency of gene expressions by their inverse frequency across all samples, aiming to highlight genes that are uniquely significant to specific diseases [7][8].

These methods convert textual gene data into structured numerical representations, facilitating the application of ML and DL algorithms for classification.

3.4 Feature Selection

Given the high dimensionality of gene expression data, feature selection is imperative to reduce feature redundancy, mitigate overfitting, and enhance model interpretability [1][3]. This study employs genetic algorithms in conjunction with the XBNet architecture to optimize feature selection. The genetic algorithms iteratively explore feature subsets, guided by a fitness function that prioritizes classification accuracy and computational efficiency [25].

Additionally, an Artificial Bee Colony (ABC) optimization technique is integrated with a modified K-Nearest Neighbors (KNN) algorithm to further refine feature selection, as demonstrated by [26]. This hybrid approach ensures the selection of the most relevant gene features, enhancing the overall performance of the classification model.

3.5 Model Training

The XBNet architecture, known for its robustness in handling high-dimensional data, is employed for model training. XBNet integrates boosted gradient descent mechanisms to optimize the learning process, thereby improving convergence rates and classification accuracy [27]. The model is trained on two versions of the dataset (V0 and V1), each subjected to both TF and TF-IDF feature extraction methods.

3.6 Evaluation Metrics

Evaluating the performance of the classification model is essential to ascertain its efficacy in predicting gene-disease associations. The following metrics are utilized:

- Accuracy: The proportion of correctly classified instances out of the total instances.

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{1}$$

OR

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Where:

TP = True Positives (correctly predicted positive cases)

TN = True Negatives (correctly predicted negative cases)

FP = False Positives (incorrectly predicted positive cases)

FN = False Negatives (incorrectly predicted negative cases)

- Precision: The ratio of true positive predictions to the total predicted positives, indicating the `model's accuracy in identifying positive cases.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Where:

TP = True Positives (correctly predicted positive cases)

FP = False Positives (incorrectly predicted positive cases)

- Recall (Sensitivity): The ratio of true positive predictions to the total actual positives, reflecting the model's ability to capture all relevant cases.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Where:

TP = True Positives (correctly predicted positive cases)

FN = False Negatives (actual positive cases that were incorrectly predicted as negative)

- F1-Score: The harmonic mean of Precision and Recall, providing a balanced measure of the model's accuracy.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \quad (5)$$

- Positive Predictive Value (PPV) and Negative Predictive Value (NPV): These metrics assess the probability that positive and negative predictions are accurate, respectively.

$$Positive\ Predictive\ Value\ (PPV) = \frac{TP}{TP + FP} \quad (6)$$

Where:

TP = True Positives (correctly predicted positive cases)

FP = False Positives (incorrectly predicted positive cases)

$$Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN} \quad (7)$$

Where:

TN = True Negatives (correctly predicted negative cases)

FN = False Negatives (actual positive cases incorrectly predicted as negative)

These metrics provide a comprehensive evaluation of the model's performance, particularly in the context of medical diagnostics where the implications of false positives and false negatives are significant [2][4].

3.7 Implementation

The implementation of the proposed methodology leverages Python, a versatile programming language widely used in data science and machine learning. The following libraries are utilized:

- NumPy: Facilitates efficient numerical computations and handling of multi-dimensional arrays, forming the backbone of scientific computing in Python [3][4].
- Pandas: Provides data structures and functions for data manipulation and analysis, enabling seamless handling of structured datasets [3][4].
- Scikit-learn (Sklearn): Offers a comprehensive suite of machine learning algorithms and tools for data preprocessing, model training, and evaluation [3][4].
- Additional Libraries: Libraries such as Matplotlib and Seaborn are employed for data visualization, aiding in the interpretation of results and model performance [3][4].

The implementation workflow involves loading the dataset, performing preprocessing and feature extraction, selecting optimal features using genetic algorithms, training the XBNet model, and evaluating its performance using the aforementioned metrics.

4. EVALUATION METRICS

In machine learning, model testing and improvement are essential. Depending on the nature of the problem, we may use a wide range of indicators to evaluate the model's effectiveness. This thesis focuses on supervised classification problems in medical imaging rather than regression metrics like mean squared error or mean absolute error.

Accurate: When categorizing data, accuracy of classification is primary metric. How well the evaluated model identifies its predictions can be simply described in this manner.

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

Given that the ratio is typically expressed as a percentage, it is not surprising that we seek a high degree of accuracy in our models. The term "binary accuracy" is used to describe this metric, and it typically results in a "good" and "bad" class being established when there are only two possible outcomes. Under these settings, four possible possibilities emerge.

A True Positive (TP) occurs when a sample from the positive class is indeed recognized as such. If this sample was incorrectly categorized as negative, it would be considered a False Negative (FN). A sample that has been accurately detected as negative is called a True Negative (TN), while a sample that has been mistakenly identified as positive is called a False Positive (FP).

Researchers in ML often use the confusion matrix [14] to visualize these metrics. Now that we have a shared vocabulary, we can think about trying a new approach to checking whether or not binary representations are accurate.

$$Binary\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

In the field of medicine, where a wrong diagnosis can have life-threatening consequences for the patient, pinpoint precision is of the utmost importance. Although doctors shouldn't place all their faith on automated diagnostic tools, high-quality data from these systems gives them an opportunity to evaluate these systems as a second opinion.

A. Predictive Values

Having established the types of outcomes that classification can result in, we further specify the following metrics:

$$Positive\ Predictive\ Value\ (PPV) = \frac{TP}{TP + FP} \tag{6}$$

Where:

TP = True Positives (correctly predicted positive cases)

FP = False Positives (incorrectly predicted positive cases)

$$Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN} \tag{7}$$

Where:

TN = True Negatives (correctly predicted negative cases)

FN = False Negatives (actual positive cases incorrectly predicted as negative)

The PPV shows how closely a "true" effect coincides with a "positive" outcome in the hypothesis test. When a patient's diagnostic tests come back positive, this probability is often used to characterize how likely it is that the patient actually suffers from the condition in question. However, the likelihood that the patient does not have this ailment is reflected by the NPV. Both the positive and negative predictive values are heavily influenced by the disease incidence rate in the study population. A higher PPV and lower NPV are the direct effect of a higher disease prevalence.

Precision: The accuracy of our model is measured by the proportion of correctly identified positives relative to all negatives. The calculations for accuracy are as follows.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

The F-score: a metric for evaluating how well a model fits a given dataset; also known as the F1-score. Binary classification systems, which divide data into "positive" and "negative" buckets, are evaluated with this metric. Harmonic mean of precision and recall is the formula for the typical F1-score. The F-score of a perfect model is 1.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \tag{5}$$

B. Implementation

Raw data is transformed into a form that can be read and analyzed by computers and machine learning algorithms during the data preparation phase of the data mining and analysis process.

Unprocessed real-world data, including text, images, and videos, is chaotic. The lack of a regular, uniform design and the presence of potential flaws are further issues.

In our model we used this libraries:

Python's import keyword allows code from one module to be used in another. Imports play a critical role in maintaining a well-organized Python program. Imports allow you to easily reuse code and keep your projects under control, both of which boost your efficiency.

a. Numpy (Numerical Python) is a free, open-source Python library used extensively across the scientific and engineering spectrum. It's the backbone of the PyData and scientific Python ecosystems and the gold standard for working with numerical data in Python. Everyone from inexperienced coders to seasoned researchers conducting cutting-edge scientific and industrial development use NumPy. The NumPy API is heavily utilized by the vast majority of other Python data science and scientific packages, such as Pandas, SciPy, Matplotlib, scikit-learn, and scikit-image. NumPy is a library that provides data structures for working with multidimensional arrays and matrices; we'll learn more about these in the following sections. It provides methods for efficiently working with the object ndarray, which is an n-dimensional array of identical elements. NumPy allows users to perform a wide variety of mathematical operations on arrays. It offers a comprehensive collection of high-level mathematical operations on such arrays and matrices, as well as robust data structures to guarantee speedy calculations.

b. Pandas: The import pandas as pd statement loads a popular Python-based data analysis package named pandas. From processing a wide range of file formats to converting an entire data table into a NumPy matrix array, it has you covered. Therefore, pandas is a trustworthy collaborator in the fields of data science and machine learning. Pandas, like NumPy, primarily deals with information stored in 1-D and 2-D arrays, though it does so in slightly different ways.

c. Sklearn: A free machine learning library for Python is called Scikit-learn. It may be used for both personal and business purposes and is a highly helpful tool for data mining and analysis. Users may execute a variety of machine learning tasks with the help of

Python Scikit-learn, which also offers a way to implement machine learning in Python. It must be compatible with Python's scientific and numerical libraries, referred known as Python NumPy and SciPy, respectively. In essence, it is a SciPy toolbox with a number of machine learning methods. We don't need to download any large standard datasets from other websites because Scikit-learn comes with modest ones. These datasets are easily importable from Python Scikit-learn.

C. Feature selection

a. Dataset V0

For the data set V0 we start with the TF feature selection, the data set divided the into two part learning and testing. The table below represent the evaluation metrics and the predictive parameters we obtained total accuracy 99%. As shown in the table 1.

TABLE I. RESULTS FOR LEARNING STEP FOR TF WITH V0

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.99	0.99	0.99	99
Asthma	1.00	1.00	1.00	
Cancer	0.99	0.98	0.99	
Diabetic	0.99	0.99	0.99	
Fabry	0.96	1.00	0.98	
Down Syndrome	1.00	0.96	0.98	

For the testing method we obtained the table bellow and the value of total accuracy is 97%. As shown in the table 2

TABLE II. RESULTS FOR LEARNING STEP FOR TF WITH V0

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.98	1.00	0.99	97
Asthma	1.00	1.00	1.00	
Cancer	0.99	0.97	0.98	
Diabetic	0.96	0.96	0.96	
Fabry	0.97	1.00	0.99	
Down Syndrome	0.95	0.93	0.94	

In the figure (1) we present the performance of the xbenet function of epoch

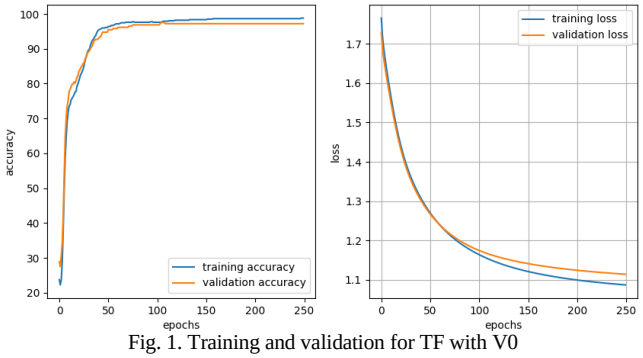

Fig. 1. Training and validation for TF with V0

In the second time we used the TF-IDF feature selection the results presented in the table below we obtain a total accuracy 31% for learning and 30 for testing. As shown in the table 3 and table 4.

TABLE III. RESULTS FOR LEARNING STEP FOR TF-IDF WITH V0

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.00	0.00	0.00	31
Asthma	1.00	0.71	0.83	
Cancer	0.22	1.00	0.36	
Diabetic	0.00	0.00	0.00	
Fabry	0.00	0.00	0.00	
Down Syndrome	0.00	0.00	0.00	

TABLE IV. RESULTS FOR TRAINING STEP FOR TF-IDF WITH V0

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.00	0.00	0.00	30
Asthma	1.00	0.46	0.63	
Cancer	0.26	1.00	0.41	
Diabetic	0.00	0.00	0.00	
Fabry	0.00	0.00	0.00	
Down Syndrome	0.00	0.00	0.00	

In the figure (2) we present the accuracy of the model function the number of the epoch.

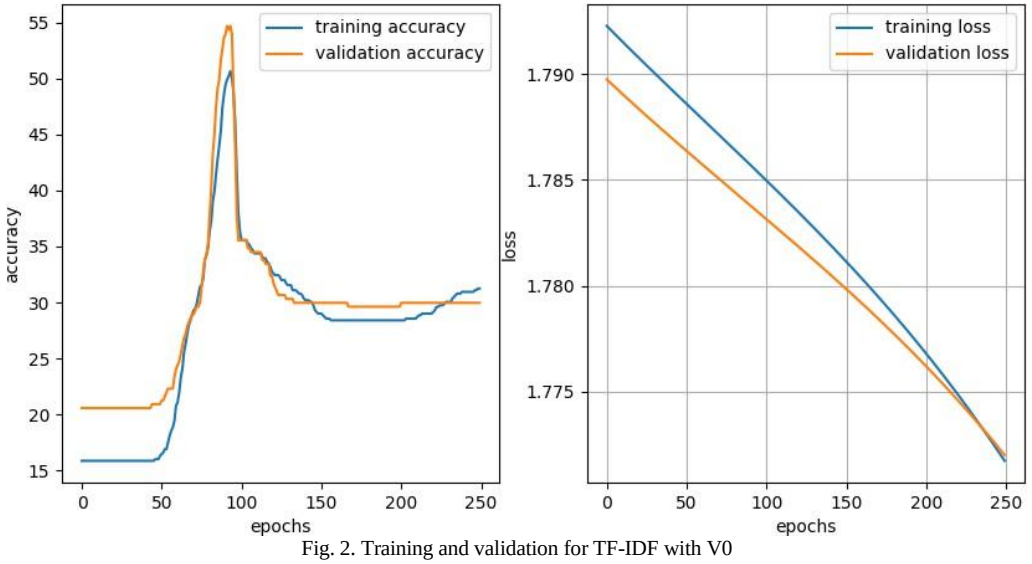

Fig. 2. Training and validation for TF-IDF with V0

b. Dataset V1

For the data set V1 we repeat the same steps,. The table below represent the evaluation metrics and the predictive parameters we obtained total accuracy 98%. As shown in the table 5.

TABLE V. RESULTS FOR LEARNING STEP FOR TF WITH V1

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.99	0.97	0.98	98
Asthma	1.00	1.00	1.00	
Cancer	0.98	0.98	0.98	
Diabetic	1.00	0.96	0.98	
Fabry	0.94	1.00	0.97	
Down Syndrome	0.99	0.98	0.99	

For the testing method we obtained the table bellow and the value of total accuracy is 98%. As shown in the table 6.

TABLE VI. RESULTS FOR TESTING STEP FOR TF WITH V1

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.98	0.98	0.98	98
Asthma	1.00	1.00	1.00	
Cancer	0.99	0.97	0.98	
Diabetic	0.98	0.98	0.98	
Fabry	0.97	1.00	0.99	
Down Syndrome	0.95	0.95	0.95	

In the figure (3) we present the performance of the algorithm XBnet in term of accuracy and loss function of epoch

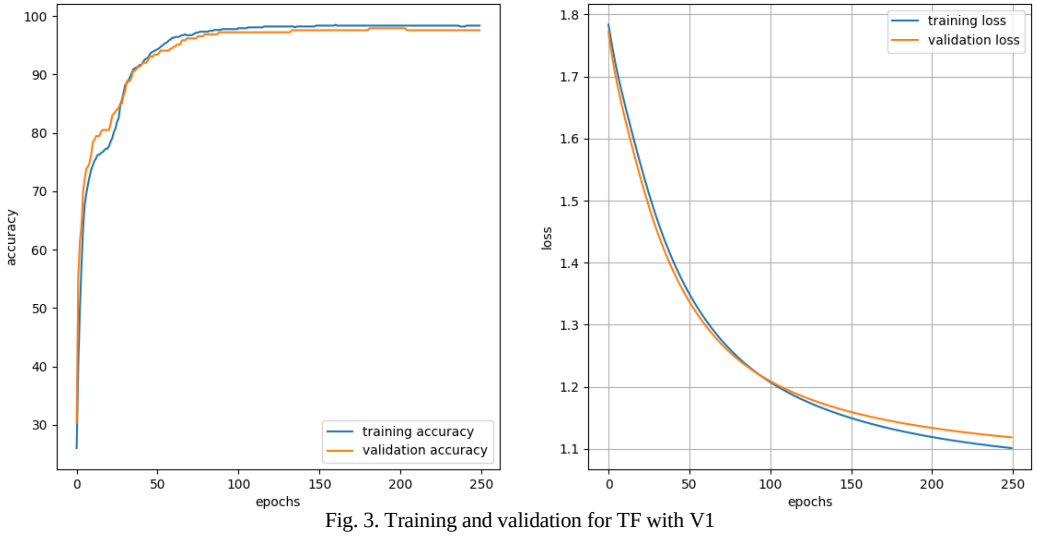

Fig. 3. Training and validation for TF with V1

In the second time we used the TF-IDF feature selection the results presented in the table below we obtain a total accuracy 20% for learning and 26% for testing. As shown in the table 7 and table 8.

TABLE VII. RESULTS FOR LEARNING STEP FOR TF-IDF WITH V1

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.00	0.00	0.00	20
Asthma	1.00	0.08	0.15	
Cancer	0.19	1.00	0.32	
Diabetic	0.00	0.00	0.00	
Fabry	0.00	0.00	0.00	
Down Syndrome	0.00	0.00	0.00	

TABLE VIII. RESULTS FOR TRAINING STEP FOR TF-IDF WITH V1

Class	Precession	Recall	F-score	Accuracy
Alzheimer	0.00	0.00	0.00	26
Asthma	1.00	0.11	0.21	
Cancer	0.25	1.00	0.40	
Diabetic	0.00	0.00	0.00	
Fabry	0.00	0.00	0.00	
Down Syndrome	0.00	0.00	0.00	

In the figure (5) we present the accuracy of the model function the number of the epoch.

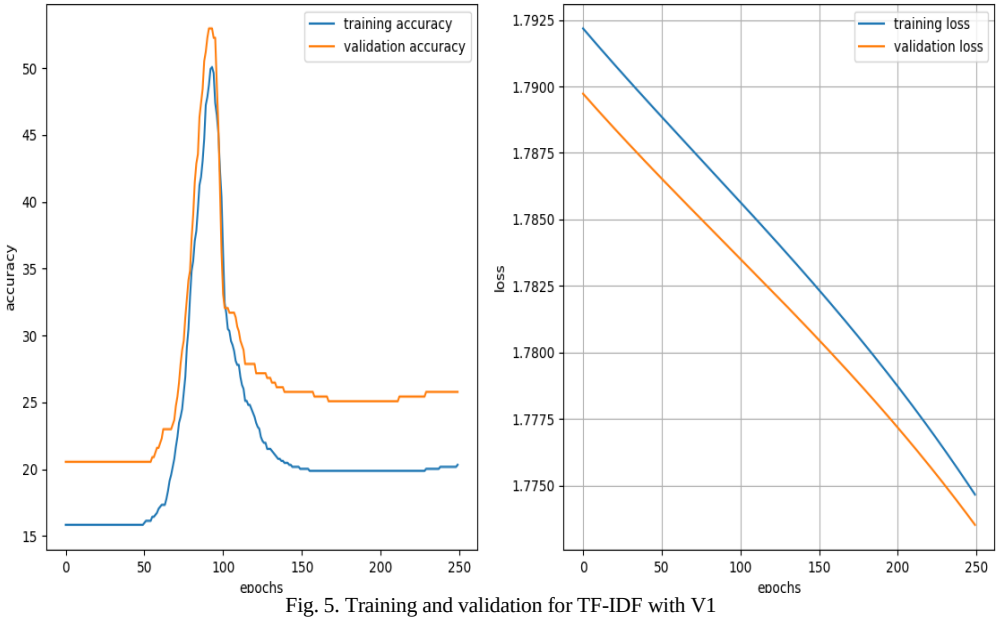

Fig. 5. Training and validation for TF-IDF with V1

From the results obtain in this thesis we can observed that with the two data set the accuracy is 98% and 97% for training and testing for feature selection used TF and for feature selection with TF-IDF the accuracy is very low the value is 31 and 26. The evaluation of the Xbnet for the we can remark that the difference between the two data set and feature selection is large. This difference due to the data set and feature selection.

The integration of genetic algorithms played a crucial role in optimizing feature selection, effectively reducing feature redundancy and enhancing the model's generalization capabilities. This synergy between optimization algorithms and advanced machine learning models significantly contributed to the high performance observed in gene-disease classification tasks.

5. CONCLUSION

In addition to basic and common used methods of feature extraction selection of genes, with a sturdy deep learning architecture we constructed this methodological framework that combines advanced feature extraction and selection techniques into one. Our method offers potential improvements in gene-disease association classification by integrating advanced feature extraction techniques with a risk open mouth that will dually serve to improve classification performance. Addressing the high-dimensionality of genomic data as well as optimization algorithms at the same time, this study aims to achieve better classification accuracy and reliability. This will become a foundation of individualized medical care and tailored therapeutic strategies.

Conflicts Of Interest

No competing financial interests are reported in the author's paper.

Funding

The absence of any funding statements or disclosures in the paper suggests that the author had no institutional or sponsor backing.

Acknowledgment

I would like to sincerely thank my collaborator, Mohammed Aljanabi, for his invaluable contributions to this research. I also wish to express my gratitude to Kabale University for their continued support and for providing the resources that made this work possible.

References

[1] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan, "Machine learning in drug discovery: A review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 1947–1999, Mar. 2022, doi: 10.1007/s10462-021-10058-4.

[2] S. R. Piccolo, A. Mecham, N. P. Golightly, J. L. Johnson, and D. B. Miller, "The ability to classify patients based on gene-expression data varies by algorithm and performance metric," *PLoS Comput. Biol.*, vol. 18, no. 3, Mar. 2022, doi: 10.1371/journal.pcbi.1009926.

[3] V. Gokulakrishnan, K. Madhubala, R. Selvasarathi, and R. Dhivya, "Microarray based disease prediction using deep learning techniques," *Int. J. Adv. Eng. Manag. (IJAEM)*, vol. 3, p. 237, 2021, doi: 10.35629/5252-0304237242.

[4] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," *Nat. Rev. Mol. Cell Biol.*, vol. 23, no. 1, pp. 40–55, Jan. 2022, doi: 10.1038/s41580-021-00407-0.

[5] C. G. Skarpathiotaki and K. E. Psannis, "Cross-industry process standardization for text analytics," *Big Data Res.*, vol. 27, Feb. 2022, doi: 10.1016/j.bdr.2021.100274.

[6] R. Egger and E. Gokce, "Natural language processing (NLP): An introduction," in *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, R. Egger, Ed. Cham: Springer Int. Publishing, 2022, pp. 307–334, doi: 10.1007/978-3-030-88389-8_15.

[7] V. Ramanathan and T. Meyyappan, "Survey of text mining," in *Proc. Int. Conf. Technol. Bus. Manage.*, Mar. 2013, pp. 508–514.

[8] V. K. A. and G. Aghila, "Text mining process, techniques and tools: An overview," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 613–622, Jul.–Dec. 2010.

[9] R. Sagayam, S. Srinivasan, and S. Roshini, "A survey of text mining: Retrieval, extraction and indexing techniques," *Int. J. Comput. Eng. Res.*, vol. 2, no. 5, 2013.

[10] V. Gupta and G. Lehal, "A survey of text mining techniques and applications," *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, Aug. 2009.

[11] M. A. Hearst, "Text data mining: Issues, techniques, and the relationship to information access," in *Proc. UW/MS Workshop on Data Mining*, Jul. 1997.

[12] R. Agrawal and M. Batra, "A detailed study on text mining techniques," *IJSCE*, vol. 2, no. 6, Jan. 2013, ISSN: 2231-2307.

[13] F. N. Patel and N. R. Soni, "Text mining: A brief survey," *Int. J. Adv. Comput. Res.*, vol. 2, no. 4, Dec. 2012.

[14] R. Patel and G. Sharma, "A survey on text mining techniques," *Int. J. Eng. Comput. Sci.*, vol. 3, no. 5, pp. 5621–5625, May 2014.

[15] M. Zortea and A. Plaza, "Spatial preprocessing for endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2679–2693, Aug. 2009.

[16] S. Supriya and M. Subaji, "Intelligent-based image enhancement using direct and indirect contrast enhancement techniques: A comparative survey," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 10, no. 7, pp. 167–184, 2017.

[17] B. Braschi, P. Denny, K. A. Gray, T. E. Jones, R. Seal, S. Tweedie, B. Yates, and E. Bruford, "Genenames.org: The HGNC and VGNC resources in 2019," *Nucleic Acids Res.*, vol. 47, pp. D786–D792, 2019.

[18] A. Snyder et al., "Genetic basis for clinical response to CTLA-4 blockade in melanoma," *N. Engl. J. Med.*, vol. 371, no. 23, pp. 2189–2199, 2014.

[19] N. Alrefaai and S. Z. Alrashid, "Classification of gene expression dataset for type 1 diabetes using machine learning methods," *Bull. Electr. Eng. Inform.*, vol. 12, no. 5, pp. 2986–2992, Oct. 2023, doi: 10.11591/eei.v12i5.4815.

[20] H. Alzoubi, R. Alzubi, and N. Ramzan, "Deep learning framework for complex disease risk prediction using genomic variations," *Sensors*, vol. 23, no. 4439, 2023.
[21] T. Jo et al., "Deep learning-based identification of genetic variants: Application to Alzheimer's disease classification," *Brief. Bioinform.*, vol. 23, no. 2, Mar. 2022, doi: 10.1093/bib/bbac022.
[22] E. Qumsiyeh, L. Showe, and M. Yousef, "GediNET for discovering gene associations across diseases using knowledge-based machine learning approach," *Sci. Rep.*, vol. 12, p. 19955, 2022, doi: 10.1038/s41598-022-24421-0.
[23] M. Sethi, S. Ahuja, S. Rani, P. Bawa, and A. Zaguia, "Classification of Alzheimer's disease using Gaussian-based Bayesian parameter optimization for deep convolutional LSTM network," *Comput. Math. Methods Med.*, vol. 2021, p. 4186666, 2021, doi: 10.1155/2021/4186666.
[24] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *J. Pers. Med.*, vol. 11, no. 2, p. 61, Jan. 2021, doi: 10.3390/jpm11020061.
[25] A. Ariani and S. Samsuryadi, "Classification of kidney disease using genetic modified KNN and artificial bee colony algorithm," *Sinergi*, vol. 25, no. 2, pp. 177–184, Jun. 2021, doi: 10.22441/sinergi.2021.2.009.
[26] Y. Gurovich et al., "Identifying facial phenotypes of genetic disorders using deep learning," *Nat. Med.*, vol. 25, pp. 60–64, 2019, doi: 10.1038/s41591-018-0279-0.
[27] M. Asif, H. F. Martiniano, A. M. Vicente, and F. M. Couto, "Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology," *PLoS One*, vol. 13, no. 12, p. e0208626, Dec. 2018, doi: 10.1371/journal.pone.0208626.