EDRAAK
Vol. **(2024)**, 2024, **pp**. 49–58
ISSN: 3078-8412

Research Article

Predictive Modeling and Analysis in Genetic Diseases: A Comprehensive Review of Recent Advances

Rana Khalid Hamad[1,*]

[1] *Islamic university of Lebanon*

ARTICLE INFO

Article History
Received 12 Mar 2024
Revised: 2 May 2024
Accepted 3 Jun 2024
Published 22 Jun 2024

Keywords
Predictive Modeling,
Genetic Diseases,
Machine Learning,
Text Mining,
Deep Learning.

ABSTRACT

This paper provides a comprehensive analysis of current advancements in predictive modeling and genetic disease classification. We delve into various machine learning techniques and text mining technologies that have significantly contributed to understanding genetic disorders and extracting valuable information from vast unstructured data. Our literature review examines key studies from recent years that have utilized machine learning models, including Naive Bayes, support vector machines, and deep learning frameworks, to improve the predictive accuracy of genetic disease outcomes. This work is aimed at enhancing the framework for predicting complex diseases using advanced computational methods.

1. INTRODUCTION

This The exploration of genetic diseases and the application of text mining to extract meaningful information from vast unstructured data sets have become pivotal areas of study in modern biomedical research. The rapid accumulation of genetic data has necessitated advanced computational approaches to understand and predict genetic disorders effectively. Machine learning techniques have shown promising results in deciphering complex biological data, aiding significantly in drug discovery and disease prediction [1][3][4]. This chapter discusses the theoretical background and practical applications of these computational techniques, with a particular focus on the challenges associated with analyzing genetic information and extracting features from unstructured text.

Genetic disorders, which arise from mutations in DNA or from quantitative deficiencies in genetic material, present a complex challenge for diagnosis and treatment. The ability to classify and predict these diseases based on genetic data varies significantly by algorithm and performance metric, highlighting the need for robust computational models [2]. Concurrently, the vast majority of data available to researchers remains unstructured and is scattered across various platforms without a clear hierarchy [7]. Text mining, therefore, emerges as a critical tool, transforming raw data into structured formats that can be easily analyzed and interpreted. Techniques in natural language processing (NLP) have evolved to assist in this transformation, making it possible to uncover hidden patterns and insights that would otherwise be inaccessible [6].

The methodological advancements in text analytics have also seen cross-industry standardization efforts, aiming to enhance the reproducibility and efficiency of text mining applications across different fields [5]. These developments underscore the significance of integrating text mining with genetic research to address the few shot dilemma where models must make accurate predictions from limited data inputs, a common scenario in genetic studies where large datasets may not always be available.

By reviewing and integrating insights from the provided studies, this chapter aims to build a comprehensive understanding of the current landscape and future directions in the prediction and analysis of genetic diseases through advanced computational methods. Through this exploration, we will highlight both the capabilities and the limitations of current technologies, setting the stage for further advancements in the field.

*Corresponding author email: r.hamad8989@gmail.com
DOI: https://doi.org/10.70470/EDRAAK/2024/008

This introduction should set a strong, research-driven foundation for your research, linking the theoretical and practical aspects of genetic diseases and text mining while pointing to the importance of integrating these areas to push the boundaries of what's possible in medical research and disease prediction.

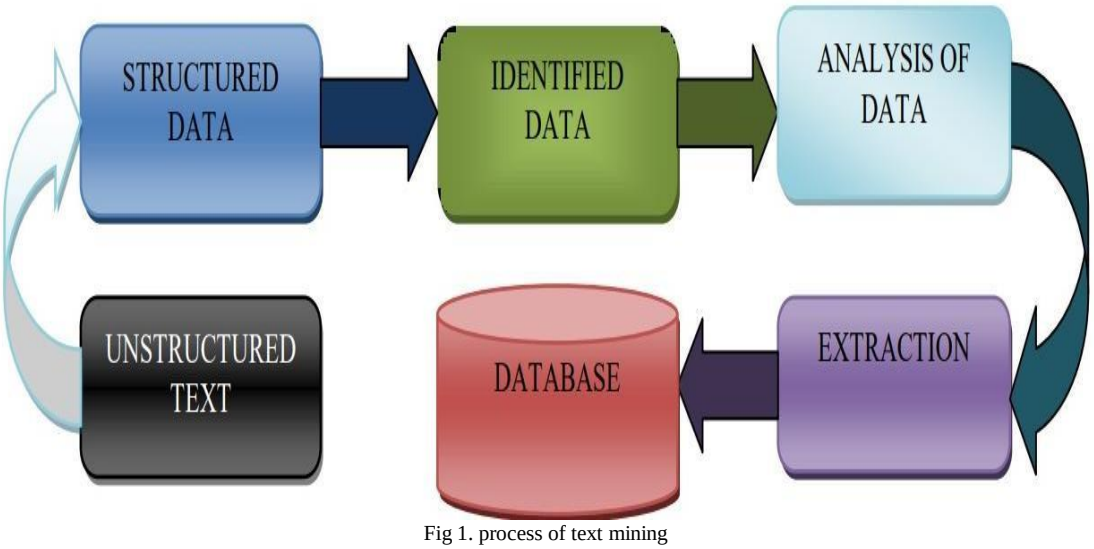

Fig 1. process of text mining

2. TEXT MINING TECHNOLOGIES

Information Retrieval: Most people think of Google when they think of information retrieval (IR) systems because of how effectively it can identify web pages on the World Wide Web that include terms that a user has provided. Document retrieval is expanded into document mining, which involves processing the documents that have been found to find the information that is useful to the user [10]. As a result, after document retrieval, there is an information extraction stage or a text summarizing stage that concentrates on the user's inquiry. In its widest sense, information retrieval (IR) includes both knowledge retrieval and information retrieval [11]. The first efforts at an automated indexing system date back to 1975, making this a reasonably established field of study. With the proliferation of the Internet and the subsequent need for sophisticated search engines, its profile grew.

Extraction Information: Extracting relevant data from text is the focus of information extraction (IE) techniques. Entity, event, and connection extraction from unstructured or semi-structured text are all highlighted. Most relevant information, including names, addresses, and companies, may be gleaned from material that hasn't been fully comprehended [12]. Meaningful data mining from texts is at the heart of IE. The term "information extraction" (IE) refers to the process of building a picture out of bits and pieces of information found in texts that are both relevant and relevant to the picture.

Categorization: A kind of "supervised" learning, text categorization requires that labels be provided for each document used in the training phase. Its primary intended use at the time was for controlled-word indexing of scholarly publications. The availability of ever- increasing amounts of digital text documents coupled with the need to arrange them for easy usage led to the field's maturation in the 1990s [13]. The process of categorization involves using the content of documents written in a standard language to place them into a set of categories. It's a group of written materials, and it's the process of determining which subject or themes best apply to each. From the standard automated or semi-automatic indexing of texts to the distribution of targeted adverts, spam filtering, the arrangement of the World Wide Web into hierarchical catalogs, the production of metadata instantly, the recognition of text genres, topic tracking, and many more purposes [14]. The study of how to automatically classify texts begins in the early 1960s. It is now a very active area of study in the science of machine learning.

Clustering: Clustering is a fascinating and crucial part of text mining. The goal is to identify hidden patterns in data and classify it into manageable chunks for in-depth research. In this method, items are sorted into clusters without any human oversight. For this challenge, you will be provided with a set of data that has not been labeled and asked to organize them into meaningful clusters. Object labels are derived entirely from the information collected. For instance, document clustering facilitates retrieval by establishing connections between related documents, facilitating the retrieval of all such documents after a single one has been determined to be relevant to a query [8]. Cluster analysis has numerous practical uses in fields as diverse as biology, pattern recognition, data mining, picture segmentation, document retrieval, business intelligence, pattern classification, security, and Web search. To accomplish data dispersion, cluster analysis may be employed independently as a text mining tool, or as a pre-processing step for other text mining algorithms to operate on the found clusters.

Summarization: Despite being an age-old problem in the field of text mining, researchers in the fields of artificial intelligence, machine learning, and NLP would do well to pay more attention to the task of summarizing texts. A text is summarized when a shorter version that nevertheless contains important information is generated automatically. When working for a large corporation, researchers often write executive summaries or highlight key aspects from papers rather than reading them in

their entirety [15]. A summary is a condensed issuance of a longer text that accurately conveys the essential points of the originals while cutting down on unnecessary detail. Technologies for text summarization include fuzzy logic, semantic graphs, decision trees, regression models, neural networks, and swarm intelligence. The quality of classifier creation varies widely and is very text-type dependent across the various approaches; this is an issue shared by all of them.

3. FEATURE EXTRACTION

Unlike texture features, which use clusters of pixels, color features only employ individual pixels. A human's visual system relies on an image's texture for both interpretation and recognition. Texture is a visual explanation for the homogeneity property of visual patterns. Two major categories of texture features (TF) are spatial TF and spectral TF. Sections 2 illustrates the numerous sub-divisions.

Spectral TF requires pictures to be translated into the frequency domain (FD) before feature extraction can take place, while spatial TF relies on pixel-level calculations in the original image. Gabor filters are often employed for TF extraction since they sample the FD of a picture by describing the orientation parameters and center frequency. The differences between these two techniques for TF extraction. Image segmentation is one application where spatial TF extraction is often employed [16]. Differences in the spatial structures of geometric or stochastic characteristics are mapped into their corresponding gray values using this method.

4. TERM FREQUENCY

Differentiating benign from malignant SPNs relied heavily on the image texture properties. Texture features may be useful to define local properties of images, quantify qualities like smoothness, roughness, and regularity, and portray particular recurring local patterns and arrangement regularity in designated image regions. The GLCM technique was used to examine the texture features of SPNs; this technique, which is based on the co-occurrence matrix model and can be used to determine the correlation between two greyscale points that are different in distance and direction, reflected the integrated information of direction, spacing, and magnitude of changes in image greyscales to describe the roughness and repeated directions of image texture.

5. TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

Many terms that are included in word vectors for clustering have no relevance to the process of extracting features from text. Although highly computationally costly, this has no impact on the accuracy of feature extraction. As a result, once we have the entire word vector, we can compute the TF-IDF for each word in the target text and then filter out any words whose TF-IDF is lower than the threshold q that we have set.

Count(w) is the number of times the word w appears in the text, and TF(w) in (1) represents this frequency. Word occurrences in the target text and in corpus samples j t are denoted by count w, whereas the total number of samples containing word w in the corpus is denoted by m.

$$TF(w) = \frac{count\ (w)}{\sum_{j=1}^{m} count(w^{t^j})} \qquad (1)$$

In equation (2), IDF(w) is the inverse file frequency of the word w, m is the number of samples in the corpus that include the word w, and n is the total number of texts in the corpus.

$$IDF\ (w) = \ln\left(\frac{n}{m+1}\right) \qquad (2)$$

Equation (3) determines the word's TF-IDF value in accordance with equations (1) and (2).

$$TFIDF(w) = \frac{n \cdot count\ (w)}{\ln\left(\frac{n}{m+1}\right) \cdot \sum_{j=1}^{m} count(w^{t_j})} \qquad (3)$$

After the low TF-IDF word is removed, equation (4) yields a set W final of all the words that are left, where Ws stands for a collection of words Ws whose TF-IDF is less than θ.

$$w_{final} = \bigcup_{j=1}^{m} w^{t^j} - w_s(TFIDF(w_s) < \theta) \qquad (4)$$

It's possible you'll want to know why we don't just train the word vector without including words with a TF-IDF below θ;

after all, that seems like it would require less work. Word2vector's mathematical process dictates that in order to train word vectors more precisely, it must take into account all words and their circumstances. The results of the training will be affected if terms that seem unrelated are left out.

Input:
R is the set of prediction rules
D is the set of documents
Output: F is the set of output file
Function: string Matching (R,D)
F= Ø
For each rule in prediction rule base R do
For each example d in D
If refers to d(R,D)
ADD information to F
Return F
Algorithm TF – IDF

6. MACHINE LEARNING

Understanding and creating "learning" methods, or methods that use data to improve performance on a certain set of tasks, is the focus of the field of study known as machine learning (ML). It's often cited as an example of AI's usefulness. Machine learning algorithms may generate predictions and judgments without being explicitly programmed to do so by building a model using sample data (also called training data). Computer vision, speech recognition, email filtering, medicine, agriculture, and other fields rely on machine learning algorithms since it is difficult or impossible to design conventional algorithms that can perform the necessary tasks.

Statistical learning is a subset of machine learning, although not all machine learning is statistical learning. Computational statistics is concerned with using computers to make predictions. Tools, theory, and application fields from the study of mathematical optimization are useful in the field of machine learning. Unsupervised learning for exploratory data analysis is the main emphasis of data mining, a related field of study. Some machine learning programs use information and neural networks in a manner that is very similar to that of the human brain. When applied to commercial issues, machine learning is sometimes referred to as predictive analytics.

7. EXTREMELY BOOSTED NEURAL NETWORK

XGBoost is a prediction-making ensemble method that employs N trees in the following way:

$$y = \psi(x) = \sum_{n=1}^{N} g_n(x) \qquad (5)$$

Feature inputs (x) and result sets (y) are denoted here. The value of the Nth leaf's score is denoted by gn(x). Additionally, gn(x) ∈M, where M is the collection of all possible ratings. Then, we apply regularization to prevent overfitting:

$$L(\psi) = \sum_i l(\hat{y}^i, y^i) + \sum_n \delta(g_n) \qquad (6)$$

where L denotes the loss function, and δ(gn) is defined as:

$$\delta(g) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_i^2 \qquad (7)$$

where the direction of regularization by λ and γ reduces overfitting. The letters T and w stand for the quantity of leaves and their relative masses. as seen in figure (2)

The extremely gradient boosted tree plays a crucial role in both our architecture and our training, and its feature importance is determined by information gain from the tree's features. This is accomplished by calculating the entropy of a set of feature vectors and then using this value to determine which attribute in the set is most useful for distinguishing between the classes to be learned. Before Li and Claramunt, there was Raileanu and Stoffel (2004). (2006). Every node's impurity level reflects how consistent the target variable is. Theoretically, we may define P to be a probability distribution that

$$P = (p_1, p_2, \ldots, p_n) \qquad (8)$$

where pi is the probability that some datapoint in D is part of some subset di In a nutshell, entropy means:

$$Entropy(P) = \sum_{i=1}^{n} -p_i \log_2(p_i) \qquad (9)$$

The information gain calculated is then used to determine the feature importance of the boosted tree which is used in

Boosted Gradient Descent

$$InformationGain = Entropy_{beforeSplit} - Entropy_{afterSplit} \quad (10)$$

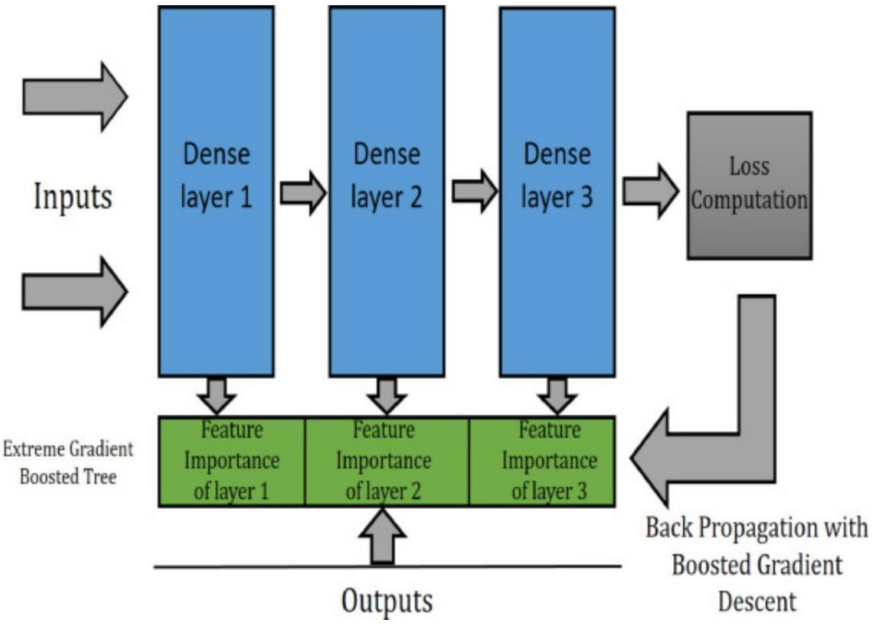

Fig 2 Xbnet Architecture

8. EVALUATION METRICS

The confusion matrix of a two-class discriminator. Predicted values are labeled as either positive (1) or negative (0), whereas actual values are denoted as True (1) or False (0). The confusion matrix contains expressions such as TP, TN, FP, and FN that can be used to make estimates of the potential of classification models[28].

- When a good result is anticipated and what really occurs is the same, we call it a True Positive (TP) in the confusion matrix. When a positive event is anticipated but a negative outcome actually occurs, the corresponding data point in the confusion matrix is considered to be false positive (FP). It is called a Type 1 Error when this happens. That's a blessing, if you're looking at it through rosy glasses of hindsight. When a negative event is anticipated but a positive outcome actually occurs, the corresponding data point in the confusion matrix is a false negative. This is a classic example of a Type 2 Error, which is just as dire a circumstance as a Type 1 Error. When a negative consequence is expected and the actual outcome is the same, the data point in the confusion matrix is True Negative (TN). These are the outcomes of the binary categorization. the proportion of correct predictions (TP + TN) to all possible predictions (P + N) is the measure of accuracy. Accuracy ranges from a peak of 1.00 to a worst of 0.00.
- Number of correct Positive Predictions (TP) discordant by the total number of positive (P) = True Positive Rate. "Recall," "Sensitivity," and "REC" are all synonyms. Maximum TP Rate is 1, minimum is 0.
- The False Positive Rate (FP Rate) is determined by dividing the number of erroneous positive predictions (FP) by the total number of negatives (N). A false positive rate of 1.0 is the worst possible rate and 0.0 is the best. In addition, 1-specificity can be used to measure it.
- Correct positive predictions (TP) are subtracted from the sum of all positive predictions (TP + FP) to arrive at the precision. Accuracy ranges from a peak of 1.0 to a bad of 0.0.
- The True Negative Rate (Specificity) is calculated by dividing the total number of accurate negative predictions (TN) by the total number of negatives (N).
  One way to evaluate a test's reliability is by looking at its F-measure or F-score. The formula used to determine this prioritizes accuracy and periodic reminders:

$$F - Score = \frac{2 \times precision \times recall}{precision \times recall} \quad (11)$$

- Cohen's Kappa coefficient (k), a measure of the number of instances classified in a machine learning model that match the data designated as the fundamental truth, controls the random classifier's accuracy as measured and anticipated accuracy.
  The Random Accuracy has a precision of 1 / k. The number of categories in the dataset is denoted by k here. Since k = 2 in a situation of binary classification, the success rate is half as high.

$$K = \frac{\boldsymbol{p0} - \boldsymbol{pe}}{(\boldsymbol{1} - \boldsymbol{pe})} \quad (12)$$

9. PREVIOUS STUDIES

In 2023, Noor AlRefaai. et al., They developed a model to classify the genes linked to type 1 diabetes using machine learning techniques. They used Naive Bayes (NB), support vector machines (SVM), and random forests (RF) to classify the genes linked to the disease in a T1D gene expression dataset with multiple classes. The model can effectively identify the T1D-related genes, which is very helpful in identifying a person who has the condition before any symptoms manifest. When using SVM with chi2 as the feature selection approach, the greatest accuracy of 89.1% was attained [19].
In 2023, Hadeel Alzoubi. et al., Develop a deep learning framework to use genetic variants to forecast the likelihood of developing complicated illnesses. A multilayer perceptron (MLP) is used in the proposed framework to predict people's state of illness. The recommended methodology was applied to the datasets from the 1958 British Birth Cohort (58C), the UK National Blood Service (NBS) Control Group, and the Welcome Trust Case-Control Consortium (WTCCC). The area under the curve (AUC) for the performance was 0.94 or higher [20].
In 2022, TaehoJo. et al., Create a deep learning-based system to identify genetic variations and apply it to the categorization of Alzheimer's disease. Propose a novel three-step approach (SWAT-CNN) for the identification of genetic variants utilizing deep learning to identify phenotype-related single nucleotide polymorphisms (SNPs) that may be utilized to develop precise disease classification models. First, the entire genome was split into optimally sized, nonoverlapping fragments. Next, convolutional neural networks (CNNs) were applied to each fragment to identify those that were related with certain phenotypes. The second phase involved running CNN on the chosen fragments using a Sliding Window Association Test (SWAT) to determine phenotypic influence scores (PIS) and discover phenotype-associated SNPs based on PIS. The final stage involved running CNN on all discovered SNPs in order to create a categorization model. GWAS data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) were used to evaluate a strategy, with N = 981, cognitively normal older people (CN) = 650, and AD = 331. method found that the most important genetic locus for AD is the well-known APOE region. Using a classification model, an area under the curve (AUC) of 0.82 was obtained [21].
In 2022, Emma Qumsiyeh. et al., Create a GediNET utilizing a knowledge-based machine learning technique to find gene connections across illnesses. created a revolutionary method called GediNET that applies past biological knowledge to gene Groups that have been linked to a particular illness, like cancer. The unique aspect of GediNET is that it afterwards makes it possible to identify meaningful connections between that particular disease and other diseases. The identification of gene Groups is the first stage in this approach. The Groups are then put through a scoring component to see which categorization Groups are doing the best. A machine learning model is then trained using the top-ranked gene groups. GediNET uses the Grouping, Scoring, and Modelling (G-S M) method to find more illnesses that share this signature in a comparable way. GediNET uses machine learning that is based on Disease- Disease Association (DDA) to find these connections. With an average of 21.61 genes, the AUC is 97% [22].
In 2021, Monika Sethi. et al., Design a Gaussian-Based Bayesian Parameter Optimized Deep Convolutional LSTM Network Classification of Alzheimer's Disease. In order to build the best deep learning model to predict the early onset of AD binary and ternary classification on MRI images, four different 2D and 3D convolutional neural network (CNN) frameworks are recommended. In addition, a few hyperparameters must be specified and changed to improve the deep learning model's performance, including learning rate, optimizers, and hidden units. Through the course of the studies, Bayesian optimization lets to utilize advantage: In addition to the findings, a persistent hyperparameter space test also provides information on the most likely outcomes. The number of experiments required to investigate space can be significantly decreased in this approach. Last but not least, long short-term memory (LSTM) through the process of augmentation has resulted in finding the better settings of the model that too in fewer iterations with a relative improvement (RI) of 7.03%, 12.19%, 10.80%, and 11.99% over the four systems optimized with manual hyperparameters tuning such that hyperparameters that look more appetizing from previous data as well as the traditional techniques of manual selection. With AD vs. MCI (S2) participants in the training and testing sets, the baseline model's accuracy is reported to be 82.65%. [23].
In 2021, Jiande Wu. et al., Create a machine learning-based classification of breast cancer types. to choose the features (genes) utilized in the construction and validation of the classification models, RNA-Sequence data from 110 triple negative and 992 non-triple negative breast cancer tumor samples from The Cancer Genome Atlas were analyzed. Support Vector Machines, K-nearest Neighbor, Naive Bayes, and Decision Tree were four distinct classification models that were assessed utilizing characteristics picked at varying threshold levels to train the models for categorizing the two forms of breast cancer. With an accuracy of 90%, a recall of 87%, and a specificty of 90%, SVM outperformed the other three classification algorithms that were tested. KNN came in second with an accuracy of 87%, a recall of 76, and a specificty of 88%. NGB and DT performed poorly on recall while being relatively accurate. The numerous characteristics employed and the imbalanced research design can help to partially explain the heterogeneity in the assessment parameters [24].
In 2021, Ardina Ariani. et al., Utilizing genetically modified knn and an artificial bee colony algorithm, classify kidney diseases. employs genetically modified K-Nearest Neighbor (KNN) and the Artificial Bee Colony (ABC) algorithm to construct a classification system for renal diseases. In order to pick the features that will best influence renal illness, the ABC algorithm is utilized. Genetically modified KNN is then used to classify the features. The three stages of study are pre-

processing, feature selection, and classification. However, it focuses on the 400 records with 24 attributes that were utilized in the pre-processing stage of chronic renal illness to pick features and classify patients. Information about kidney disease is divided into two categories: chronic kidney disease and non-chronic kidney disease. The outcome revealed that by splitting the dataset into 280 training and 120 test data, an accuracy of 98% was attained [25].

In 2019, Yaron Gurovich. et al., Create a system for employing deep learning to identify the face characteristics of hereditary diseases. provide a system for analyzing face images called Deep Gestalt that measures resemblances to a large number of syndromes using computer vision and deep learning methods. In three first studies, two with the aim of identifying people with a specific condition from other syndromes and one with the purpose of separating several genetic subtypes in Noonan syndrome, Deep Gestalt outperformed physicians. Deep Gestalt identified the right diagnosis on 502 distinct photos with top-10 accuracy in the final testing, which reflected a genuine clinical setting problem.

A community-driven phenotyping platform collected a dataset of more than 17,000 photos spanning more than200 syndromes, which was used to train the model. Clinical genetics, genetic testing, research, and precision medicine might all benefit significantly from the potential benefits of Deep Gestalt [26].

In 2018, Muhammad Asif. et al., Create a system for finding disease-related genes by comparing comparable genes functionally using the Gene Ontology. created a method for supervised machine learning to predict the genes of complex diseases. Candidate genes for the autism spectrum disorder (ASD) were used to evaluate the proposed pipeline. By using several semantic similarity metrics, a quantitative measure of gene functional similarity was discovered. Different kinds of machine learning classifiers were developed based on quantitative semantic similarity matrices of ASD and non-ASD genes in order to identify the underlying functional similarities between ASD genes. The classifiers improved upon previously reported ASD classifiers were trained and evaluated on ASD and non-ASD gene functional similarity. A Random Forest (RF) classifier outperformed the previously reported classifier (0.73) in its ability to detect novel ASD genes, with an AUC of 0.80. This classifier was also able to identify 73 unique ASD candidate genes that were enriched for fundamental ASD traits including obsessive-compulsive disorder and autism. Attention deficit hyperactivity disorder (ADHD) and other co-occurring illnesses with ASD were also elevated in predicted genes. Additionally, the suggested technique was used to create a KNIME workflow that users can customize and utilize without having to have programming or machine learning expertise [27].

TABLE I SUMMARY PREVIOUS STUDIES

No.	Name of Authors	Year	Methods	Dataset	Result
1	Noor AlRefaai. et al.,[29]	2023	Machine learning approaches by applying machine learning (ML) approaches for classification, such as Naive Bayes (NB), support vector machines (SVM), and random forests (RF).	T1D gene expression dataset includes multiclass classification of the genes linked to this condition	Accuracy of 89.1% as obtained
2	Hadeel Alzoubi. et al.,[30]	2023	employs a multilayer perceptron (MLP)	applied to the 1958 British Birth Cohort (58C) dataset, the UK National Blood Service (NBS) Control Group, and the Wellcome Trust Case-Control Consortium (WTCCC) dataset.	The performance had an area under the curve (AUC) of 0.94 or higher.
3	TaehoJo. et al.,[31]	2022	Deep learning's (SWAT-CNN) three- step method for identifying genomic variations	GWAS data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) were used, with (N= 981; CN =650; AD= 331) and cognitively normal older individuals (CN) as well as AD as the dependent variables.	Using a classification model, an area under the curve (AUC) of 0.82 was obtained.
4	Emma Qumsiyeh. et al.,[32]	2022	Use GediNET, which combines previous scientific information with gene groups that have been linked to a particular illness, such as cancer.	10 human gene expression datasets from the GEO database for	With an average of 21.61 genes, the AUC is 97%.

				various complicated disorders	
5	Monika Sethi et al.,[33]	2021	Utilizi Deep Convolutional LSTM Network Gaussian-Based Bayesian Parameter Optimization	ADNI Dataset	testing sets is observed at 82.65%
6	Jiande Wu. et al.,[34]	2021	Four different classification models including Support Vector Machines, K- nearest neighbor, Naïve Bayes and Decision tree	The Cancer Genome Atlas Program (TCGA)	SVM had the best performance with an accuracy of 90%, a recall of 87% and a specificty of 90%, followed by KNN, with an accuracy of 87%, a recall of 76 and specificty of 88%.
7	Ardina Ariani. et al.,[35]	2021	artificial bee colony algorithm and genetically altered knn	chronic dataset derived fro m the UCI Repository Machine Learning.	By splitting the dataset into 280 training and 120 test data, it was possible to achieve an accuracy of 98%, according to the results.
8	Yaron Gurovich. et al.,[36]	2019	using computer vision and deep-learning algorithms	Full London Medical Databases dataset of thousands of images	DeepGestalt achieved 91% top-10 accuracy in identifying the correct syndrome on 502 different images.
9	Muhammad Asif. et al.,[37]	2018	devised a method for supervised machine learning to forecast the genes for complicated diseases.	The Simons Foundation Autism Research Initiative (SFARI) gene database provided information on ASD.	The AUC for the Random Forest (RF) classifier was 0. 80.

10. CONCLUSION:

The studies reviewed in this paper underscore the effectiveness of machine learning and text mining in genetic disease prediction. From the use of deep learning to identify phenotypic patterns to the application of text mining for feature extraction from medical texts, the research community has made significant strides. The ongoing evolution of computational technologies promises even greater advances, potentially leading to more personalized and accurate medical interventions for genetic disorders. This paper not only highlights the current state of research but also sets the stage for future developments in the field.

Funding:

The authors did not receive any financial support, grants, or sponsorships from public, private, or not-for-profit sectors. This study was self-financed.

Conflicts of Interest:

The authors declare no conflicts of interest with the research presented.

Acknowledgment:

The authors appreciate their institutions' encouragement, advice, and provision of essential resources to facilitate this study.

Reference

[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. Manage.*, 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 International 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. Manage., 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, 2012.

[10] V. Gupta and G. Lehal, "A survey of text mining techniques and applications," J. Emerg. Technol. Web Intell., vol. 1, no. 1, pp. 1–11, Aug. 2009.

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

[12] R. Agrawal and M. Batra, "A detailed study on text mining techniques," Int. J. Softw. Comput. Eng., vol. 2, no. 6, Jan. 2013.

[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] T. 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. Zaki 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.

[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, bbac022, 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, no. 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, pp. 1–16, 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, Y. Hanani, and O. Bar, "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, V. 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, Dec. 2018, doi: 10.1371/journal.pone.0208626.

[28] X. Wang, L. Wang, and J. Bi, "Anomaly detection in wireless sensor networks using machine learning and data fusion techniques," Sensors, vol. 20, no. 6, p. 1560, 2020.

[29] N. AlRefaai et al., "Machine learning approaches for classification using Naive Bayes, support vector machines, and random forests: T1D gene expression dataset multiclass classification," 2023.

[30] H. Alzoubi et al., "Application of a multilayer perceptron (MLP) to the 1958 British Birth Cohort (58C) dataset, the UK National Blood Service (NBS) Control Group, and the Wellcome Trust Case-Control Consortium (WTCCC) dataset," 2023.

[31] T. Jo et al., "A three-step deep learning method (SWAT-CNN) for identifying genomic variations in GWAS data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)," 2022.

[32] E. Qumsiyeh et al., "GediNET: Combining prior scientific knowledge with gene groups linked to diseases such as cancer using GEO database datasets," 2022.

[33] M. Sethi et al., "Utilizing Deep Convolutional LSTM Network with Gaussian-Based Bayesian Parameter Optimization on the ADNI Dataset," 2021.

[34] J. Wu et al., "Evaluation of classification models including Support Vector Machines, K-nearest neighbor, Naïve Bayes, and Decision Tree on The Cancer Genome Atlas Program (TCGA)," 2021.

[35] A. Ariani et al., "Artificial bee colony algorithm and genetically altered KNN on a chronic dataset derived from the UCI Repository Machine Learning," 2021.

[36] Y. Gurovich et al., "Using computer vision and deep-learning algorithms on the Full London Medical Databases dataset," 2019.

[37] M. Asif et al., "Supervised machine learning method to forecast genes for complicated diseases using the SFARI gene database for ASD," 2018.