ISSN 2723-6471

An Effective Investigation of Genetic Disorder Disease Using Deep Learning Methodology

B. Vidhya[1,*], B. L. Shivakumar[2], Siti Sarah Maidin[3], Jing Sun[4]

[1]*Department of CS with Data Analytics, Sri Ramakrishna College of Arts and Science, Coimbatore-641006, India*

[2]*Department of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore-641006, India*

[3]*Faculty of Data Science and Information Technology (FDSIT), INTI International University, Nilai, Malaysia*

[4]*Faculty of Liberal Arts, Shinawatra University, Thailand*

(Received: June 22, 2024; Revised: July 26, 2024; Accepted: August 30, 2024; Available online: September 17, 2024)

Abstract

This study evaluates the performance of four neural network models—Artificial Neural Network (ANN), ANN optimized with Artificial Bee Colony (ANN-ABC), Multilayer Feedforward Neural Network (MLFNN), and Forest Deep Neural Network (FDNN)—across different iteration levels to assess their effectiveness in predictive tasks. The evaluation metrics include accuracy, precision, Area Under the Curve (AUC) values, and error rates. Results indicate that FDNN consistently outperforms the other models, achieving the highest accuracy of 99%, precision of 98%, and AUC of 99 after 250 iterations, while maintaining the lowest error rate of 2.8%. MLFNN also shows strong performance, particularly at higher iterations, with notable improvements in accuracy and precision, but does not surpass FDNN. ANN-ABC offers some improvements over the standard ANN, yet falls short compared to FDNN and MLFNN. The standard ANN model, though improving with iterations, ranks lowest in all metrics. These findings highlight FDNN's robustness and reliability, making it the most effective model for high-precision predictive tasks, while MLFNN remains a strong alternative. The study underscores the importance of model selection based on performance metrics to achieve optimal predictive accuracy and reliability.

Keywords: Genetic Disease, Deep Learning, Forest Deep Neural Network, Supervised Learning, Feature Detector, Ensemble Learning, Process Innovation, Inclusive Health

1. Introduction

Gene processing involves the identification and analysis of genes within DNA sequences [1]. Quantitative information is extracted from genomic data using various computational techniques, such as gene prediction, gene expression analysis, and gene function annotation [2], [3]. In this process, image processing techniques play a critical role, especially in analyzing microarray images, which help quantify gene expression levels. During microarray image processing, each spot on the array is identified, its intensity is measured, and it is compared against the background. This process yields gene expression data that can be used to study the dynamics of gene expression over time, across different tissues, and in relation to disease states [4], [5].

The integration of gene processing with image processing—two essential tools in genetic research—provides valuable insights into the genetic origins of numerous disorders and facilitates the development of personalized treatments [6]. However, identifying defective genes and the genetic conditions from ever-expanding genomic data remains a complex and time-consuming task. Traditional methods for disease identification based on gene analysis are labor-intensive. To overcome these challenges, several algorithms have been developed to detect defective genes and the genetic causes of diseases [7], [8]. In gene analysis, researchers often rely on ontologies to interpret and integrate data from multiple sources, enabling cross-species data analysis [9]. However, challenges related to content curation and annotation persist. To address these issues, AI-based approaches have been introduced. Artificial Intelligence (AI) involves a combination of computing frameworks, theories, and algorithms that enable machine perception, speech recognition, reasoning,

***Corresponding author: B.Vidhya (vidhya.b@srcasac.in)**
DOI: https://doi.org/10.47738/jads.v5i3.370

ISSN 2723-6471

natural language understanding, and decision-making [10], [11]. AI encompasses several techniques, including computer vision, machine learning, natural language processing, rule-based logic, and deep learning. AI-based methods can significantly accelerate the analysis of vast amounts of data, detecting patterns and producing quick results that inform further decision-making [12], [13].

By leveraging sophisticated algorithms, AI generates analytical models that uncover patterns and predict outcomes [14]. The rise of big data and the growing demand for data analysis have heightened the need for computing power, storage, and effective data processing. The insights produced by these analyses are actionable and valuable for advancing genetic research [15]. In this study, a multilayer feedforward neural network (FFN) is employed to model the potential interactions between genes. Weight predictions and signal direction are used to strengthen interaction links among genes. The proposed FFN model is validated using Monte Carlo cross-validation, which optimizes generalization ability and reduces the risk of overfitting the model. By investigating the network of genes associated with genetic diseases, the model highlights the significant role specific genes play in the genomic data and disease outcomes.

The remainder of this article is structured as follows: Section 2 presents a literature review on diseases caused by genetic disorders, Section 3 describes the proposed Forest Deep Neural Network-based disease classification, Section 4 discusses the numerical results of the proposed FFN, and Section 5 concludes the article.

2. Related Work

Andersen et al. [10] explored the structure of genes, genetic disease linkages, and pseudogenes. Their experimental investigation defined the composition of protein tyrosine phosphatases (PTP) in the human genome. By utilizing proprietary sequences and public databases, they described exon structures and chromosomal loci, predicting alternative splicing and amino acid sequences. The study linked the identification of diseased genes to the PTP sequence.

López-Bigas and Ouzounis [11] proposed a computational approach for identifying genes involved in genetic diseases in humans. Through sequence analysis, they identified protein groups associated with hereditary disease-causing genes. Cooper and Krawczak [12] studied the patterns of genes responsible for human diseases using single base pair substitutions and mutational spectrum analysis. Similarly, Cooper and Youssoufian [13] developed a method for identifying single base pair mutations involving CpG dinucleotides, aiding in the mapping of the genome and the identification of genetic diseases.

Wang et al. [14] constructed a three-dimensional model of proteins to gain significant insights into disease-causing genes. Their research focused on the molecular mechanisms of these genes, leading to the identification of specific genes responsible for human diseases. Daetwyler et al. [15] introduced a broad approach to assess the genetic risk of diseases. Their method improves the accuracy of hereditary disorder risk prediction. Veltman and Brunner [16] focused on de novo mutations, aiming to identify rare gene variations linked to complex genetic diseases.

Wray et al. [17] conducted an association study on genetic risk prediction, emphasizing the complexity of genetic disease susceptibility due to nucleotide variations in primary genes. Asif et al. [18] applied machine learning and gene ontology to identify diseased genes. Given the complexity of biological marker identification, gene ontology was employed to clarify gene function. Similarly, Deng et al. [19] integrated gene ontology with multiple networks to identify microRNAs, which play a key role in human biological processes. Inferring the function of microRNAs is crucial in disease identification.

Schlicker et al. [20] prioritized the semantic similarity of gene ontologies to improve disease prediction, while Ortutay and Vihinen [21] combined protein interaction networks and gene ontologies to identify candidate genes associated with disease. Mohammadi et al. [22] used gene ontology and microarray data mining to detect disease-causing genes. Le and Dang [23] highlighted the challenge of identifying biomarkers in complex hereditary diseases due to their multifactorial etiology, proposing the use of network similarity to identify diseased genes.

Vidaki et al. [24] applied artificial neural networks (ANN) to predict age based on DNA methylation, a technique valuable in forensic science. Garro et al. [25] utilized ANN and Artificial Bee Colony optimization algorithms to classify DNA microarrays, allowing for the prediction, approximation, and classification of genomic data.

 ISSN 2723-6471

Atkov et al. [26] integrated clinical parameters and genetic polymorphism with ANN to identify heart diseases, while Coppedè et al. [27] applied ANN to assess the risk of Down syndrome by detecting chromosome damage and polymorphisms. Khan et al. [28] used ANN for gene expression profiling, helping to predict and classify genes linked to human cancers. Finally, Koçer and Canal [29] combined genetic algorithms with ANN to classify epilepsy. They noted that predicting disease through ANN is complex, especially when biomarkers for certain genes are challenging to assign.

3. Proposed Methodology

The feature selection and classification of diseases resulting from genetic disorders are covered in this section.The process of feature selection is attained by feature learning and the classification is attained by Forest Deep Neural Network (FDNN). The proposed methodology combines the strengths of Random Forest (RF) and Deep Neural Network (DNN) into an FDNN model. In this approach, the DNN acts as a learner, and the forest component functions as a feature detector using training data to predict outcomes with novel feature representations. The forest section consists of several independent decision trees, each generating a binary result. These binary results are merged and converted into a one-hot encoding, which is then fed into the DNN for further learning [30]. Various algorithms can be used to construct the forest component, but in this study, the RF algorithm is employed to identify features from raw inputs.

For the Small Round Blue Cell Tumors (SRBCTs) cDNA microarray data, the input vectors are presented in a 699 x 9 matrix format. This matrix consists of 699 input vectors, with each column representing one of nine average values, calculated as 4.42, 3.14, 3.21, 2.81, 3.22, 3.46, 3.44, 2.87, and 1.59, respectively. To ensure that each input component aligns within the learning range of the FDNN, normalization constants a=0.1a = 0.1a=0.1 and b=0.8b = 0.8b=0.8 are applied. This transformation scales the input components to fall within the range [0.1, 0.9], with the following normalized average values: 0.41, 0.29, 0.3, 0.26, 0.3, 0.38, 0.32, 0.27, and 0.15.

The normalized data, comprising eight input components and their corresponding target outputs, are divided into training and testing sets for further analysis. The FDNN model consists of two main components: the forest section, which acts as a feature detector, learning sparse representations from raw inputs under the supervision of the training data, and the DNN section, which predicts outcomes based on these learned feature representations [31]. The forest is formed by constructing independent decision trees, and it operates as an ensemble of these trees. As the Random Forest algorithm is ideal for constructing the forest component, it has been used in this study. However, other forest structures can be explored depending on the nature of the feature space. For example, network-guided forests may be employed if the feature space is well-structured, or simple bagging techniques may be used to build the forest when feature space is unknown.

The architecture of the FDNN is illustrated in figure 1. The FDNN classifier training process consists of two phases. In the first phase, the forest is trained using labeled training data, and in the second phase, a fully-connected DNN is trained using the predictions from each tree in the forest for every instance [32]. After completing the two-stage training, the trained forest and DNN work together to compute predictions for test instances across the entire model.

As shown in figure 1, the forest prediction feature f_i, $\forall_i$ is one-hot encoded for implementation purposes. Since the final output dimension of the DNN is two, this operation is analogous to encoding label vectors y_i. Therefore, the input to the DNN in this implementation is represented as a tensor with dimensions n×M×2, rather than a matrix of n×M. In the DNN model, the activation function used is the rectified linear unit (ReLU), which has an advantage over sigmoid and hyperbolic tangent functions by avoiding the vanishing gradient problem during optimization. The method is implemented in Python using the Scikit-learn and TensorFlow libraries.

The proposed mathematical approach for constructing the FDNN successfully classifies 387 out of 388 training data samples, achieving a classification accuracy of 99.74%. The generalization performance of the model can be summarized as follows: out of 311 test data, the model correctly classifies all 311 samples (100%). Overall, FDNN demonstrates an exceptional performance, correctly identifying 698 out of 699 data points, resulting in an accuracy of 99.86%.

The trained FDNN model is further utilized to analyze and score the associations between genes, specifically measuring the strength of the relationships between source and target genes, which may either be suppressive or stimulatory. This analysis allows for the construction of an association map for the genes, with positive or negative signs indicating

 ISSN 2723-6471

stimulatory or suppressive interactions, respectively. To filter out the least relevant interactions, the Pearson correlation coefficient rrr is employed with a cutoff value of 0.7.

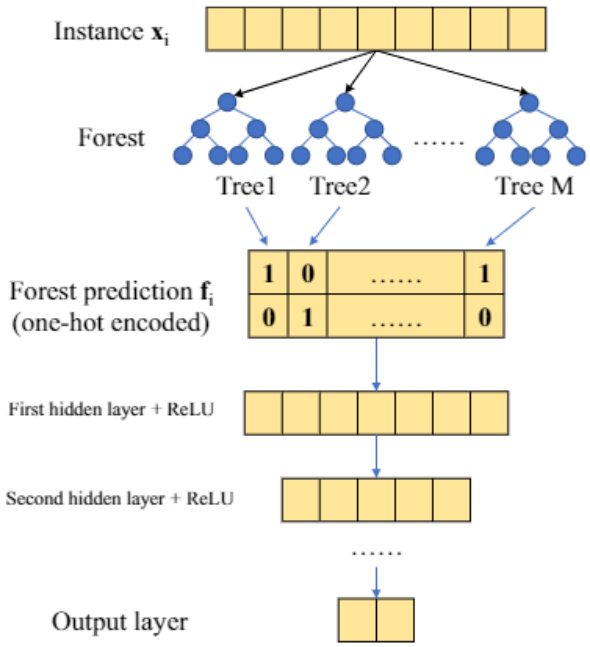

Figure 1. Architecture of FDNN Model

4. Results and Discussion

This section discusses about the dataset description and performance metrics. Additionally, the numerical outcome of the proposed approach is illustrated with comparative analysis. The performance of the proposed approach is compared with ANN ABC based ANN, MLFNN with DFNN.

4.1. Dataset Description

Khan et al. [36] conducted an investigation on the cDNA microarray dataset of SRBCTs with the aim of identifying marker genes capable of distinguishing between four distinct forms of blue cell tumors frequently confused in pediatric cases. Employing Artificial Neural Network (ANN) classification models and Principal Component Analysis (PCA), the study utilized an initial dataset comprising 88 samples, each characterized by 2,308 genes. These samples were categorized into four tumor types: rhabdomyosarcoma (RMS), Ewing's sarcoma (EWS), Burkitt lymphoma (BL), and neuroblastoma (NB). Specifically, the distribution included 25 samples in the RMS group, 29 in the EWS group, 11 in the BL group, 18 in the NB group, and the remaining 5 samples were categorized as unknown. The objective was to discern distinctive genetic markers contributing to accurate differentiation among these tumor types with the potential to enhance diagnostic precision in pediatric cases.

4.2. Performance Metrics

Accuracy:The degree to which the value derived from classified occurrences is close to the genuine value is measured by accuracy. It represents both enduring defects and quantitative bias and includes true positive (TP) and true negative (TN) values for each of the evaluated classes. The least accuracy is associated with a discrepancy between the result and actual values. It refers specifically to the identification of instances relative to the actual value. The calculation of accuracy is delineated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

Precision:The closeness of the observation and the significance of the results found are indicated by the precision, or positive quantitative value. The proportionate rate of p Accuracy and precision are interchangeable synonyms. False Positive (FP) and True Positive (TP) rates are used in its calculation. Precision is correlated with the percentage of positive values in the total population. The amount of actual positive attributes (i.e., the count of the item successfully identified as positive classes) represents the accuracy estimate for a particular issue in the classification stage. Consequently, the very precise algorithm generates more needed data than extraneous data.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

Gain:The reduction of entropy or surprise value by transformation of a dataset that is frequently used for training is known as information gain. It is calculated by,

$$\text{Gain} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$

Area Under the Curve (AUC):AUC serves as a thorough synopsis of the Receiver Operating Characteristic (ROC) curve and offers a measurable assessment of a classifier's class discrimination performance. An increased AUC value indicates that the model is more capable of differentiating between positive and negative classifications.

4.3. Result

Table 1 presents a comparison of accuracy across different models—ANN, Artificial Neural Network optimized with Artificial Bee Colony (ANN-ABC), Multilayer Feedforward Neural Network (MLFNN), and FDNN—over a series of iterations. The number of iterations used to train each model ranges from 50 to 250. The results indicate that the FDNN model consistently outperforms the other models at every iteration level. Specifically, after 50 iterations, FDNN achieves an accuracy of 92.5%, which increases steadily to 99% after 250 iterations. In comparison, MLFNN begins with an accuracy of 90.5% at 50 iterations and reaches 95% by 200 iterations, maintaining the same level of accuracy at 250 iterations. The ANN-ABC model shows a slight improvement over the standard ANN model, starting with 86% accuracy at 50 iterations and reaching 91.5% at 250 iterations. Meanwhile, the standard ANN model achieves the lowest accuracy overall, starting at 83% after 50 iterations and increasing gradually to 89% after 250 iterations.

Table 1. Comparison of Accuracy

Iterations	ANN	ANN-ABC	MLFNN	FDNN
50	83	86	90.5	92.5
100	84	88	92	94
150	87	90.5	93	95
200	88	91	95	98
250	89	91 .5	95	99

As depicted in figure 2, FDNN (yellow line) consistently outperforms the other models in terms of accuracy, starting at around 92.5% after 50 iterations and progressively increasing to 99% at 250 iterations. MLFNN (gray line) shows steady accuracy improvement, starting at 90.5% after 50 iterations and reaching 95% by 200 iterations, maintaining this level at 250 iterations. ANN-ABC (orange line) shows a moderate improvement over the standard ANN (blue line) throughout the iterations. ANN-ABC starts at 86% accuracy at 50 iterations and reaches 91.5% at 250 iterations, while the standard ANN model begins at 83% accuracy and gradually climbs to 89% by the end of 250 iterations.

ISSN 2723-6471

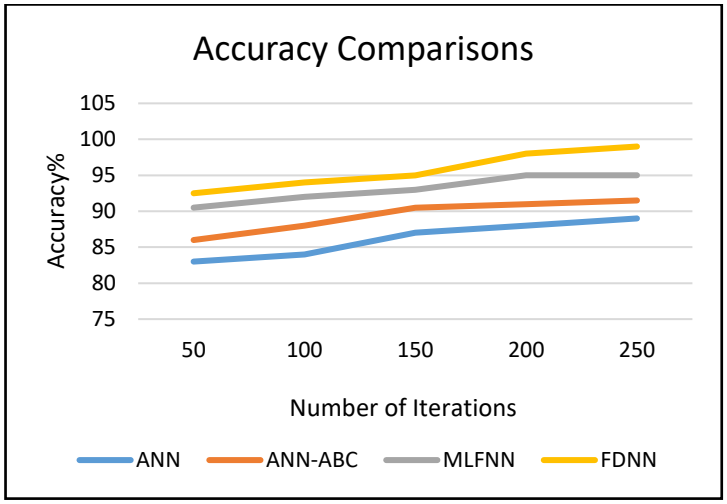

Figure 2. Comparison of Accuracy

Table 2 display the precision of various models across different numbers of iterations. The models compared include the ANN, ANN-ABC, MLFNN, and FDNN. Precision is measured in percentage (%). At 50 iterations, FDNN achieves the highest precision at 90%, followed by MLFNN at 89%, ANN-ABC at 84%, and ANN at 77.5%. With 100 iterations, FDNN maintains its lead with a precision of 92%, while MLFNN increases to 90.5%, ANN-ABC rises to 84.5%, and ANN slightly improves to 78%. At 150 iterations, FDNN reaches its peak precision of 94%, with MLFNN at 91%, ANN-ABC at 86%, and ANN at 79%. By 200 iterations, FDNN's precision advances to 95%, MLFNN reaches 92.5%, ANN-ABC achieves 86.5%, and ANN stabilizes at 79%. Finally, at 250 iterations, FDNN shows the highest precision of 98%, followed by MLFNN at 94%, ANN-ABC at 88%, and ANN at 82%. This table highlights that FDNN consistently demonstrates superior precision across all tested iterations, showing a steady improvement. MLFNN also exhibits significant precision improvement but does not surpass FDNN. ANN-ABC performs better than ANN but falls short compared to MLFNN and FDNN. ANN shows the lowest precision overall, although it does improve with additional iterations. This analysis provides valuable insights into the efficiency and effectiveness of the different models as the number of iterations increases.

Table 2. Comparison of Precision

Iterations	ANN	ANN-ABC	MLFNN	FDNN
50	77.5	84	89	90
100	78	84.5	90.5	92
150	79	86	91	94
200	79	86.5	92.5	95
250	82	88	94	98

Table 3 presents a comparison of the area under the curve (AUC) values for various models across different iteration levels. The models compared include ANN, ANN-ABC, MLFNN, and FDNN. At 50 iterations, FDNN achieved the highest AUC of 94, followed by MLFNN at 91, ANN-ABC at 82, and ANN at 78.5. With 100 iterations, FDNN maintained the top position with an AUC of 96, while MLFNN recorded 92, ANN-ABC 83, and ANN 80.5. After 150 iterations, FDNN further improved its performance to an AUC of 98, with MLFNN at 94, ANN-ABC at 86, and ANN at 82. At 200 iterations, FDNN reached its peak AUC of 98.5, surpassing MLFNN (95.5), ANN-ABC (86.5), and ANN (83.5). Finally, at 250 iterations, FDNN achieved the highest AUC of 99, demonstrating superior performance compared to MLFNN (96), ANN-ABC (87), and ANN (85). This table highlights that FDNN consistently delivers the highest AUC values across all iteration levels, indicating its robust performance compared to other models. MLFNN also shows significant improvements, especially at higher iterations, making it a strong performer as well. In contrast, while ANN and ANN-ABC improve with additional iterations, they do not reach the performance levels of MLFNN and FDNN.

ISSN 2723-6471

Table 3. Comparison of AUC

Iterations	ANN	ANN-ABC	MLFNN	FDNN
50	78.5	82	91	94
100	80.5	83	92	96
150	82	86	94	98
200	83.5	86.5	95.5	98.5
250	85	87	96	99

Table 4 presents a comparison of the error rates for various models at different iteration levels. The models evaluated are ANN, ANN-ABC, MLFNN, and FDNN. At 50 iterations, FDNN exhibits the lowest error rate of 2.1, outperforming MLFNN (4.2), ANN-ABC (6.3), and ANN (7.44). As the number of iterations increases to 100, FDNN continues to show the lowest error rate of 2.2, with MLFNN at 4.3, ANN-ABC at 6.2, and ANN at 7.66. At 150 iterations, FDNN's error rate increases slightly to 2.3, while MLFNN's rate is 4.4, ANN-ABC's is 6.4, and ANN's is 7.8. By 200 iterations, FDNN's error rate rises to 2.5, compared to MLFNN at 4.8, ANN-ABC at 6.6, and ANN at 8.3. At 250 iterations, FDNN still maintains the lowest error rate of 2.8, with MLFNN at 4.9, ANN-ABC at 6.9, and ANN at 8.4. This table highlights that FDNN consistently achieves the lowest error rates across all iteration levels, demonstrating its superior accuracy relative to other models. While error rates for ANN, ANN-ABC, and MLFNN increase with more iterations, FDNN remains the most accurate model throughout, reflecting its robust performance in minimizing errors.

Table 4. Comparison of Error Rate

Iteration	ANN	ANN-ABC	MLFNN	FDNN
50	7.44	6.3	4.2	2.1
100	7.66	6.2	4.3	2.2
150	7.8	6.4	4.4	2.3
200	8.3	6.6	4.8	2.5
250	8.4	6.9	4.9	2.8

4.4. Discussion

The results presented in table 1, table 2, table 3, and table 4 provide a comprehensive evaluation of the performance of various models—ANN, ANN optimized with Artificial Bee Colony (ANN-ABC), Multilayer Feedforward Neural Network (MLFNN), and FDNN—across different iteration levels. The findings reveal clear distinctions in model performance regarding accuracy, precision, AUC values, and error rates, with FDNN consistently outperforming the other models. As illustrated in table 1 and figure 2, FDNN demonstrates superior accuracy across all iteration levels, starting at 92.5% after 50 iterations and achieving a remarkable 99% accuracy after 250 iterations. This consistent improvement in accuracy underscores FDNN's effectiveness and robustness in handling the task. MLFNN also shows notable accuracy improvements, reaching a plateau of 95% by 200 iterations and maintaining this level at 250 iterations. ANN-ABC shows a moderate performance improvement compared to the standard ANN, but it does not match the accuracy levels of FDNN or MLFNN. The standard ANN model consistently lags behind, starting with the lowest accuracy and showing the slowest improvement over iterations.

Table 2 reveals that FDNN leads in precision at all tested iterations. Its precision increases from 90% at 50 iterations to 98% at 250 iterations, further validating its overall superior performance. MLFNN also shows significant precision gains but remains below FDNN. ANN-ABC, while outperforming the standard ANN, does not achieve the precision levels of FDNN and MLFNN. The standard ANN model consistently shows the lowest precision, though it does improve slightly with more iterations. Table 3 highlights FDNN's dominance in AUC values. Starting with an AUC of 94 at 50 iterations and reaching 99 at 250 iterations, FDNN consistently delivers the highest AUC, reflecting its superior ability to discriminate between classes. MLFNN performs well, especially at higher iterations, but does not surpass FDNN. ANN and ANN-ABC, although improving with additional iterations, lag behind FDNN and MLFNN in AUC

performance. As shown in table 4, FDNN maintains the lowest error rates throughout all iteration levels. Its error rate, although slightly increasing with iterations, remains the lowest compared to MLFNN, ANN-ABC, and ANN. This indicates that FDNN not only achieves higher accuracy and precision but also effectively minimizes errors. The increasing error rates of ANN, ANN-ABC, and MLFNN with additional iterations suggest that FDNN's performance in error minimization is more consistent and reliable.

The results consistently indicate that FDNN is the most effective model among those tested, exhibiting superior performance in accuracy, precision, AUC values, and error rates. MLFNN also performs well, particularly at higher iterations, but does not surpass FDNN. While ANN-ABC offers some improvements over the standard ANN, it falls short in comparison to FDNN and MLFNN. Overall, these findings highlight FDNN's robust capability and effectiveness in handling the task at hand, making it the preferred model for achieving optimal performance.

5. Conclusion

This study provides a thorough evaluation of several neural network models—ANN, ANN-ABC, MLFNN, and FDNN—across various iteration levels. The results demonstrate clear performance differences, with FDNN emerging as the most effective model in terms of accuracy, precision, AUC values, and error rates. FDNN consistently outperforms the other models, achieving the highest accuracy of 99% and the highest precision of 98% after 250 iterations. It also maintains the highest AUC value of 99 and exhibits the lowest error rate of 2.8%, highlighting its robust capability in delivering accurate and reliable predictions. The superior performance of FDNN across all metrics indicates its effectiveness in handling complex tasks and suggests that it is a highly reliable model for applications requiring high precision and minimal error rates.

MLFNN also shows significant performance improvements, particularly at higher iterations, but does not reach the levels achieved by FDNN. MLFNN demonstrates strong accuracy, precision, and AUC values, positioning it as a strong alternative to FDNN. ANN-ABC, while improving over the standard ANN, does not achieve the performance levels of FDNN or MLFNN. The standard ANN model consistently ranks lowest in accuracy, precision, and AUC values, and exhibits higher error rates, making it less effective compared to the other models. FDNN stands out as the most efficient and reliable model among those tested, making it the preferred choice for tasks requiring high accuracy and precision. MLFNN, despite its strong performance, does not surpass FDNN but remains a viable option. ANN-ABC and standard ANN show limited improvements and are less effective in comparison. These findings underscore the importance of selecting the appropriate model based on performance metrics to achieve optimal results in predictive tasks.

6. Declarations

6.1. Author Contributions

Conceptualization: B.V., B.L.S., S.S.M., dan J.S.; Methodology: B.L.S., S.S.M., dan J.S.; Software: B.V.; Validation: B.V., B.L.S., S.S.M., dan J.S.; Formal Analysis: B.V., B.L.S., S.S.M., dan J.S.; Investigation: B.V.; Resources: S.S.M.; Data Curation: S.S.M.; Writing Original Draft Preparation: B.V., B.L.S., S.S.M., dan J.S.; Writing Review and Editing: S.S.M. dan B.V.; Visualization: B.V. dan J.S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 1020-1030, 2018.

[2] S. G. Tangye, W. Al-Herz, A. Bousfiha, T. Chatila, C. Cunningham-Rundles, A. Etzioni, L. C. Notarangelo, B. Puck, A. S. Sullivan, and K. E. Sullivan, "Human inborn errors of immunity: 2019 update on the classification from the International Union of Immunological Societies Expert Committee," *Journal of Clinical Immunology*, vol. 40, no. 1, pp. 24-64, 2020.

[3] P. A. Flume, J. D. Chalmers, and K. N. Olivier, "Advances in bronchiectasis: endotyping, genetics, microbiome, and disease heterogeneity," *The Lancet*, vol. 392, no. 10150, pp. 1247-1259, 2018.

[4] M. A. Kelly, C. Caleshu, A. Morales, J. Buchan, Z. Wolf, S. M. Harrison, and B. Funke, "Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel," *Genetics in Medicine*, vol. 20, no. 3, pp. 252-260, 2018.

[5] C. Has, J. W. Bauer, C. Bodemer, M. C. Bolling, L. Bruckner-Tuderman, A. Diem, and L. M. Lindhout, "Reclassification of inherited epidermolysis bullosa and other disorders with skin fragility," *British Journal of Dermatology*, vol. 182, no. 6, pp. 1474-1482, 2020.

[6] A. Morales, J. Buchan, Z. Wolf, S. M. Harrison, and B. Funke, "Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel," *Genetics in Medicine*, vol. 20, no. 3, pp. 252-260, 2018.

[7] M. A. Haendel, C. G. Chute, and P. N. Robinson, "Classification, ontology, and precision medicine," *New England Journal of Medicine*, vol. 379, no. 14, pp. 1354-1360, 2018.

[8] E. Cornec-Le Gall, V. E. Torres, and P. C. Harris, "Genetic complexity of autosomal dominant polycystic kidney and liver diseases," *Journal of the American Society of Nephrology*, vol. 29, no. 2, pp. 492-508, 2018.

[9] C. F. Wright, B. West, M. Tuke, S. E. Jones, K. Patel, T. W. Laver, and M. N. Weedon, "Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting," *The American Journal of Human Genetics*, vol. 104, no. 1, pp. 54-65, 2019.

[10] J. N. Andersen, P. G. Jansen, M. E. Peters, M. J. H. Robinson, and S. S. Zhang, "Genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage," *The FASEB Journal*, vol. 22, no. 3, pp. 739-748, 2008.

[11] N. López-Bigas and C. A. Ouzounis, "Genome-wide identification of genes likely to be involved in human genetic disease," *Nucleic Acids Research*, vol. 32, no. 11, pp. 3344-3352, 2004.

[12] D. N. Cooper and M. Krawczak, "The mutational spectrum of single base-pair substitutions causing human genetic disease, patterns and predictions," *Human Genetics*, vol. 86, no. 4, pp. 119-129, 1990.

[13] D. N. Cooper and H. Youssoufian, "The CpG dinucleotide and human genetic disease," *Human Genetics*, vol. 93, no. 1, pp. 6-16, 1998.

[14] X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, and H. Yu, "Three-dimensional reconstruction of protein networks provides insight into human genetic disease," *Nature Biotechnology*, vol. 30, no. 3, pp. 240-244, 2012.

[15] H. D. Daetwyler, B. Villanueva, and J. A. Woolliams, "Accuracy of predicting the genetic risk of disease using a genome-wide approach," *PLoS ONE*, vol. 3, no. 10, pp. 1-10, 2008.

[16] J. A. Veltman and H. G. Brunner, "De novo mutations in human genetic disease," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 565-575, 2012.

[17] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk to disease from genome-wide association studies," *Genome Research*, vol. 17, no. 11, pp. 1520-1528, 2007.

[18] M. Asif, H. F. Martiniano, A. M. Vicente, and F. Couto, "Identifying disease genes using machine learning and gene

functional similarities, assessed through Gene Ontology," *PLoS ONE*, vol. 13, no. 5, pp. 1-12, 2018.

[19] L. Deng, J. Wang, and J. Zhang, "Predicting gene ontology function of human microRNAs by integrating multiple networks," *Journal of Computational Biology*, vol. 26, no. 4, pp. 321-332, 2019.

[20] A. Schlicker, T. Lengauer, and M. Albrecht, "Improving disease gene prioritization using the semantic similarity of Gene Ontology terms," *Bioinformatics*, vol. 26, no. 18, pp. 2202-2208, 2010.

[21] C. Ortutay and M. Vihinen, "Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies," *Nucleic Acids Research*, vol. 47, no. 1, pp. 91-104, 2019.

[22] A. Mohammadi, M. H. Saraee, and M. Salehi, "Identification of disease-causing genes using microarray data mining and Gene Ontology," *BMC Medical Genomics*, vol. 4, no. 1, pp. 1-10, 2011.

[23] D. H. Le and V. T. Dang, "Ontology-based disease similarity network for disease gene prediction," *Vietnam Journal of Computer Science*, vol. 3, no. 2, pp. 73-82, 2016.

[24] A. Vidaki, D. Ballard, A. Aliferi, T. H. Miller, L. P. Barron, and D. S. Court, "DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing," *Forensic Science International: Genetics*, vol. 29, no. 1, pp. 196-203, 2017.

[25] B. A. Garro, K. Rodríguez, and R. A. Vázquez, "Classification of DNA microarrays using artificial neural networks and ABC algorithm," *Applied Soft Computing*, vol. 38, no. 1, pp. 485-496, 2016.

[26] O. Y. Atkov, S. G. Gorokhova, A. G. Sboev, E. V. Generozov, E. V. Muraseyeva, S. Y. Moroshkina, and N. N. Cherniy, "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters," *Journal of Cardiology*, vol. 59, no. 2, pp. 183-192, 2012.

[27] F. Coppedè, E. Grossi, F. Migheli, and P. Migliore, "Polymorphisms in folate-metabolizing genes, chromosome damage and risk of Down syndrome in Italian women: identification of key factors using artificial neural networks," *BMC Medical Genomics*, vol. 3, no. 1, pp. 37, 2010.

[28] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.

[29] S. Koçer and M. R. Canal, "Classifying epilepsy diseases using artificial neural networks and genetic algorithm," *Journal of Medical Systems*, vol. 35, no. 3, pp. 545-554, 2011.

[30] B. Srinivasan and T. Wahyuningsih, "Navigating Financial Transactions in the Metaverse: Risk Analysis, Anomaly Detection, and Regulatory Implications," *Int. J. Res. Metav.*, vol. 1, no. 1, pp. 59-76, 2024.

[31] B. H. Hayadi and I. M. M. El Emary, "Enhancing Security and Efficiency in Decentralized Smart Applications through Blockchain Machine Learning Integration", *J. Curr. Res. Blockchain.*, vol. 1, no. 2, pp. 139–154, Sep. 2024.

[32] Henderi and Q. Siddique, "Comparative Analysis of Sentiment Classification Techniques on Flipkart Product Reviews: A Study Using Logistic Regression, SVC, Random Forest, and Gradient Boosting,"*J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 21-42, 2024.