THE SCIENTIFIC TEMPER (2025) Vol. 16 (12): 5355-5364

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.23

https://scientifictemper.com/

RESEARCH ARTICLE

Optimized Hybrid Feature Selection Techniques for Detecting Iron Deficiency Anemia

S. Srinithiya[1], K. Menaka[2*]

Abstract

The Iron Deficiency Anemia (IDA) is one of the most common types of nutritional disorders in the world and it requires precise and timely diagnosis to avoid the consequences of its development in the human body. This work aims is to improve and boost the classification performance of diagnosing IDA by utilizing different Feature Selection Techniques (FST) on the basis of filter, wrapper, embedded and hybrid approaches. A dataset containing the biological markers was compiled for analysis and several algorithms like Analysis of Variance (ANOVA) F-statistic, Recursive Feature Elimination (RFE), Least Absolute Shrinkage and Selection Operator (LASSO), Mean Squared Error (MSE), Random Forest and Support Vector Machine (SVM) from the above FST were used to determine the most discriminative features. Also, some hybrid algorithms from statistical and model-based selection, including ANOVA with Logistic Regression (Anolog) and Random Forest with Chi-square (ChiForest) were developed and evaluated. Based on their performance, the most valuable features were selected and thus the performance evaluation is enhanced. This comprehensive study highlights the effectiveness of hybrid feature selection methods to enhance the diagnostic accuracy, the model efficiency and clarity of interpretation. It is suggested by the findings that advanced machine learning and feature selection techniques should be integrated to come up with robust diagnostic tools that could be used to identify IDA.

Keywords: Iron Deficiency Anemia(IDA), Feature Selection Techniques(FST), Filter, wrapper and Embedded methods, Hybrid feature selection techniques.

Introduction

In the modern data-driven world of healthcare, a substantial part of the global population suffers due to a variety of health disorders and anemia, especially Iron Deficiency Anemia (IDA), continues to be one of the most common and under-diagnosed health conditions. Healthcare institutions and diagnostic laboratories have stored large quantities of medical informatics, but it is difficult to analyze this information and extracting meaningful insights can be challenging without the use of intelligent tools. The use of artificial intelligence (AI), specifically machine learning (ML) has enabled healthcare researchers and practitioners to develop the way of detecting, classifying and predicting a disease in a range of fields including cardiovascular diseases, cancer and nutritional deficiencies (Thomas Davenport et al., 2019).

As high-dimensional datasets multiply, it is becoming more complicated to identify the relevant patterns and discover the crucial biomarkers. This is especially severe in hematological conditions, where the presence of overlapping symptoms may lead to confusion and misinterpretation in diagnosis. The system must identify feature selection as an important pre-processing element to remove the amount of redundant or irrelevant features and make the model effective and interpretable. When it comes to the diagnosis of anemia, selecting the main hematological parameters, such as hemoglobin (HGB), RBC, ferritin, WBC, folate, MCV and vitamin B12 deficiency anemia is crucial. Some of the methods used in optimization include feature selection methods such as filter, wrapper, embedded and hybrid techniques (Senthil Murugan Nagarajan et al., 2021).

[1]Research Scholar, *PG & Research Department of Computer Science,* UrumuDhanalakshmi College, *(Affiliated to Bharathidasan University),* Tiruchirappalli, Tamil Nadu, India.

[2]Head & Assistant Professor, *PG & Research Department of Computer Science,* UrumuDhanalakshmi College, *(Affiliated to Bharathidasan University),* Tiruchirappalli, Tamil Nadu, India.

***Corresponding Author:** K. Menaka, Optimized Hybrid Feature Selection Techniques for Detecting Iron Deficiency Anemia, E-Mail: k.menaka@udc.ac.in

How to cite this article: Srinithiya, S., Menaka, K. (2025). Optimized Hybrid Feature Selection Techniques for Detecting Iron Deficiency Anemia. The Scientific Temper, **16**(12):5355-5364.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.23

Source of support: Nil

Conflict of interest: None.

© The Scientific Temper. 2025

Received: 24/11/2025 **Accepted:** 09/12/2025 **Published:** 25/12/2025

The performance of feature selection techniques is that they increase the accuracy of classification and decrease the model complexity and execution time. The filter methods provide computational efficiency through the selection of features on the basis of statistical significance and the wrapper methods iteratively evaluate feature subsets and assessing their performance on a classifier. The embedded techniques make use of feature selection as part of the model-building process (Beatriz Remeseiro et al., 2019). Hybrid methods are the combination of two or more methods and it possess the ways to increase predictive reliability and to balance time consumption.

This study compared various feature selection techniques with standardized measures of performance metrics to assess the progress towards reliable, interpretable, and high-performance diagnostic frameworks in the field of hematology. After this process, the algorithms which performed excellently good among all the metrics were chosen and two hybrid algorithms were designed among them to get still more better results. The outcome of the results helps to detect IDA at an early stage and improved diagnostic precision, promoting better treatment strategy for anemia patients.

Literature Review

N. Thomas et al., 2020, conducted a survey of feature selection methods which is emphasized as a basic preprocessing task that can help machine learning by eliminating dimensionality and avoid overfitting and increase interpretability. The common FST methods are systematically reviewed and their respective merits in terms of predictive power and computational speed are identified. The survey underlines that embedded techniques are combined with random forests or support vector machines, may perform better than the other feature selection techniques. Based on their research the best methods of feature selection promote not only quality prediction but also efficient development of high statistical and interpretative performance.

M.A.Shanti et al., 2024, compared the traditional filter based algorithm including Information Gain and Chi-square with the wrapper-based and hybrid approaches to assess the efficacy of the methods in improving model accuracy, cost of computation and the ability to handle high dimensionality. In healthcare, various medical conditions are being used to perform the evaluation of the performance based on different classification measurements such as accuracy, precision, recall, and F1-score. An investigation was carried out into the robustness and scalability of the methods when dealing with high-dimensional datasets, through which significant performance differences were revealed.

H. Ayyıldız et al., 2020, explored the use of machine learning that can help to distinguish between iron deficiency anemia (IDA) and beta-thalassemia based on red blood cell parameters. They used the Neighborhood Component Analysis (NCA) feature selection technique to discover the most relevant hematological features that can be used to classify correctly. In their study, they established the fact that machine learning models, especially when trained with NCA, effectively increase the accuracy of diagnosing similar hematologic diseases. The results indicate the opportunity of combining computational methods and clinical diagnostics to facilitate better medical decisions.

Another work recommended by Terzi et al., 2022, developed a novel expert system based on Iron Deficiency Anemia (IDA) diagnosis through artificial intelligence methodology. The system integrates the data and utilizes it based on the use of the rule-based logic to assist with making precise medical decisions. Their method showed great results in terms of diagnostic accuracy and reliability, which reduced the necessity of excessive testing. Their survey focuses on the reliability of the intelligent systems to automate the process of early detection and better clinical outcomes in the diagnosis of anemia.

Siddartha Pullakhandam et al., 2024, proposed the classification and interpretation of IDA with Complete Blood Count (CBC) data in a machine learning-based solution. To enhance accuracy in diagnosis and clinical decision-making, the researchers used a wide range of algorithms employed by the ML to achieve accuracy. It also aimed at explainability by making sure healthcare professionals can comprehend the models predictions. The findings demonstrated that ML approaches can be used as an efficient method to automate the classification of anemia with transparency of the feature contribution.

Asare et al., 2024, suggested a new method of machine learning to diagnose Iron Deficiency Anemia (IDA) in children based on conjunctiva images. This study has utilized the preprocessing, feature extraction and classification of the images to reliably detect without invasive tests. The findings showed a high level of accuracy, highlighting the potential of computer vision with respect to screenings of anemia in pediatric patients.

M.D.Dithy et al., 2020, proposed an anemia screening framework for pregnant women, in which feature selection was combined with data classification algorithms. The most relevant hematological parameters for classification were identified, leading to enhanced diagnostic accuracy. Various machine learning models were tested and performance was improved while computational complexity was reduced through the use of feature selection. This approach was demonstrated to have potential for early detection and improved management of anemia during pregnancy.

Tounsi S et al., 2022, indicated that the diagnosis of breast cancer was improved through the use of feature selection methods incorporated into machine learning. In their research, several approaches were interrelated to determine the most applicable predictors for classification.

Better detection was revealed when classifier models were trained on features refined through feature selection methods. In the study, the benefit of using targeted feature selection to enhance diagnostic accuracy without altering computational efficiency was emphasized.

Jaar Abdollahi et al., 2022, developed a hybrid model that combined feature selection and the ensemble classifier method in the optimization of heart disease diagnosis. A combination of the filter and wrapper methods was used to identify the most relevant clinical features, followed by training of multiple classifiers whose classification results were stacked. This type of multilayered approach proved to perform much better in terms of diagnostics with models. The combination of feature selection and ensemble learning was also noted as a way of creating robust and accurate efficient predictive systems in medical diagnostics.

Chaganti et al., 2022, presented a work for thyroid disease prediction using selective feature selection combined with machine learning. An optimized Subset of biomarkers was selected based on statistical and relevance filters and the features used to train different classification models. The fact that models trained on this set of carefully selected features achieved high predictive accuracy showed that reducing feature selection was successfully performed without undermining predictive accuracy.

Vinnarasi et al., 2025, suggested two new hybrid techniques of feature selection CorrRecursive Feature Selection (CRFS) and RanChi Ensemble Selection (RCES). In their paper, the advantages of the filter and wrapper approach have been merged together to find out the features that were strongly related to the levels of TSH and Vitamin D. Both hybrid techniques were highly effective with regard to accuracy, sensitivity, specificity, and F-measure compared to traditional filter, wrapper, and embedded methods. The results showed the value of hybrid feature selection approaches in enhancing predictive performance and physiological interactions.

Pathan et al., 2022, reported that the role of feature selection in enhancing heart disease prediction was investigated using machine learning models on high-dimensional clinical datasets. It was demonstrated that the use of relevant feature subsets significantly improved classification accuracy while reducing model complexity and time. Despite the reduction in the number of input features, high model performance, highlighting the efficiency gains achievable through proper feature selection.

Methods and Materials

System Framework

The goal of this work is to optimize classification accuracy by selecting the most relevant features from the dataset and reducing redundancy. The framework integrates critical steps like data gathering, preprocessing, feature selection strategies, training of classifiers, and performance assessment of the models. This work emphasizes the importance of feature selection methods in improving the accurate categorization of Iron Deficiency Anemia.

Data Acquisition

The dataset employed in this study was derived from the open-source dataset (Kaggle, 2023) shared in the Kaggle repository which consists of blood based biological parameters that are routinely used for the diagnosis of anemia. It consists of 5201 patient records with 30 features, collected from healthcare settings. The selection of these biomarkers is fundamental because they provide valuable diagnostic insight into IDA as well as different nutritional anemia types. The attributes cover both numerical features (e.g., WBC, HGB, MCV, Ferritin, Folate, Vitamin B12, etc) and nominal features (e.g., gender, anemia class labels). The dataset also contains class labels, such as *Iron_anemia_Class, Folate_anemia_class, B12_anemia_class* and a target variable *(class)* as shown in Table 1. This dataset provides a comprehensive basis for applying feature selection techniques to improve the accurate classification of Iron Deficiency Anemia, thereby enhancing diagnostic precision.

Data Preprocessing

In this stage, the raw dataset is systematically examined, refined and transformed into a suitable format for analysis and modelling. This section will make sure the features being considered have been appropriately scaled and normalized to enhance the effectiveness of the methods of feature selection and aid in the stability of machine learning models.

Data Loading

The pre-processed data were then stored in a Comma Separated Values (CSV) file format which is highly compatible with machine learning applications. The dataset is loaded with the help of the Python environment with Jupyther notebook which offers effective tools to work with tabular data. The .csv dataset file was read into a Pandas DataFrame using the read_csv() function. This enabled the data to be organized under a structured tabular format, with each row containing a patient record and each column containing haematological features.

Handling Missing Values

A frequent problem in datasets or incomplete records is missing values. Mishandling of missing values may adversely impact model results and give biased inferences. Therefore, the dataset was analyzed thoroughly concerning the data missing prior to using machine learning algorithms. The dataset was imported into a Pandas DataFrame, and the functions like isnull()were applied to missing values in all the 30 features.

Table 1: Description of the IDA Dataset

Feature Name	*Feature Type*	*Description*
GENDER · WBC · NE# · LY# · MO#	Nominal	Gender of the patient
WBC	Numerical	White Blood Cell count indicator of immune response
NE#	Numerical	Neutrophil count important in bacterial infection defense
LY#	Numerical	Lymphocyte count related to immune function
MO#	Numerical	Monocyte count contributes to immune regulation
EO#	Numerical	Eosinophil count elevated in allergies and parasitic infections
BA#	Numerical	Basophil count related to inflammatory responses
RBC	Numerical	Red Blood Cell count indicates oxygen-carrying capacity
HGB	Numerical	Hemoglobin concentration used to diagnose anemia
HCT	Numerical	Hematocrit proportion of blood volume occupied by RBCs
MCV	Numerical	Mean Corpuscular Volume, average size of red blood cells
MCH	Numerical	Mean Corpuscular Hemoglobin, average hemoglobin per RBC
MCHC	Numerical	Mean Corpuscular Hemoglobin Concentration, hemoglobin concentration in RBCs
RDW	Numerical	Red Cell Distribution Width , measures variation in RBC size
PLT	Numerical	Platelet count, important for blood clotting
MPV	Numerical	Mean Platelet Volume, average size of platelets
PCT	Numerical	Plateletcrit, volume percentage of platelets in blood
PDW	Numerical	Platelet Distribution Width, variation in platelet size
SD	Numerical	Standard Deviation of RBC size, related to anisocytosis
SDTSD	Numerical	Standard deviation measure linked with RBC/platelet indices
TSD	Numerical	Total Standard Deviation, statistical variation in measurements
FERRITTE	Numerical	Serum Ferritin, indicator of iron storage in the body
FOLATE	Numerical	Serum Folate, important for DNA synthesis and RBC production
B12	Numerical	Vitamin B12 level , essential for RBC maturation
All_Class	Nominal	Combined anemia class label (multiple deficiencies)
HGB_Anemia_Class	Nominal	Anemia classification based on Hemoglobin levels
Iron_anemia_Class	Nominal	Label indicating Iron Deficiency Anemia (IDA)
Folate_anemia_class	Nominal	Label indicating Folate Deficiency Anemia
B12_Anemia_class	Nominal	Label indicating Vitamin B12 Deficiency Anemia
Class	Nominal	Target, Final class label

Data Inspection

The info() command is used to give basic information on the dataset. This provides details about the dataset format like the number of records, data type and presence of missing data.

Splitting of Data

The data were split into training and testing sets to make sure that the machine learning models could be generalized. The correct division of data is crucial to avoid excessive overfitting and to estimate objectively model performance on unseen data.

Training–Testing Split

The train_test-split Scikit-learn function that uses 80:20 split ratio between training and testing (John Smith et al., 2021). A stratified division was made according to the target variable to maintain the distribution of anemic and non-anemic cases in both sets, which is important to maintain the balance in classes during evaluation.

Feature Selection

The importance of feature selection (FS) is to improve the diagnostic accuracy of Iron Deficiency Anemia (IDA). Since the dataset includes the hematological features (e.g., RBC,

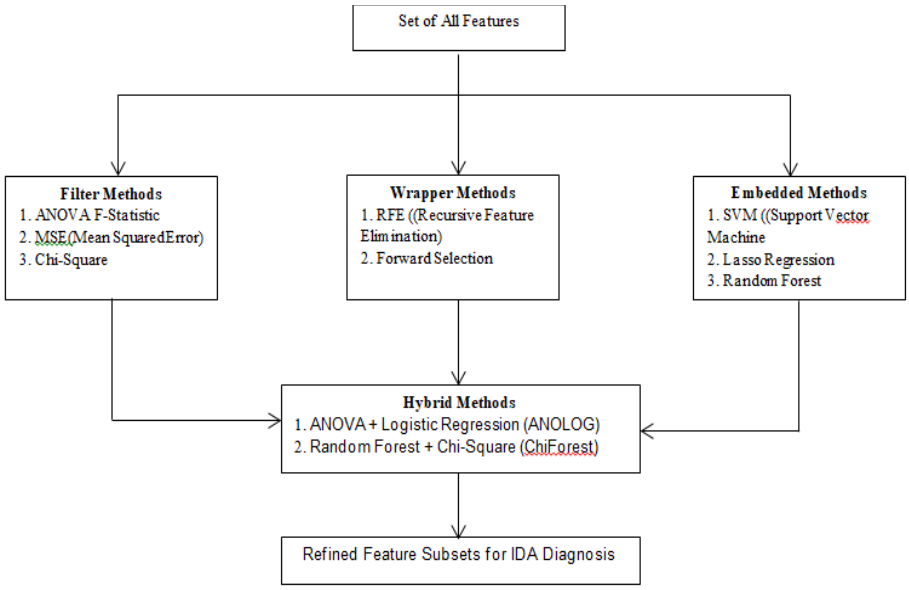

Figure 1: System framework for feature selection

MCV, HGB, Ferritin, Folate, B12, etc.) not every feature will be equally helpful in identifying the IDA. Redundant or irrelevant variables may reduce the quality of the model, add to computational expense and make interpretation more difficult. This stage can be used to make models that are precise and effective, only through the identification and retention of the most informative features. This process plays a critical role in minimizing the complexity of training and consequently enhancing the general efficiency of learning. The selection of features was done systematically in this study by using a variety of existing feature selection techniques, including the newly proposed hybrid strategies. Fig.1 demonstrates the use of Filter, Wrapper, Embedded, and Hybrid methods to refine feature selection for the detection of IDA. These resulting subsets are the most informative attributes according to the respective selection mechanisms. Each technique or combination of techniques are intended to be used in feature selection to improve the model performance to minimize overfitting and enhance interpretability (Vohra et al., 2022).

Filter methods

Filter methods are used as a preprocessing step, with the most informative features being identified before a machine learning algorithm is employed. This method is aimed at determining the most informative features based on the statistical assessment of their relationship with the target variable. Ranking is followed by the retention of only the highest scoring features to build the model. These approaches are computationally effective since they are independent on a specific learning algorithm and are particularly appropriate in high-dimensional data. The most important step is to eliminate irrelevant or weakly related properties at the first step and streamline the process in feature selection methods (Moorthy et al., 2020).

Although a variety of filter methods were examined, two of them yielded more encouraging outcomes than the others. These techniques are:

- ANOVA F- Statistic
- Mean squared error (MSE)
- Chi-Square

Wrapper methods

This feature selection approach works by evaluating different subsets of features and choosing the subset that produces the best performance. Wrapper methods provide a powerful, though resource-intensive, approach to feature selection that balances interpretability with strong predictive performance. In healthcare, these techniques can be used to dramatically increase accuracy in classification and also to reduce the number of features needed to detect and decision support systems (T Saw et al., 2019).

Some of the methods which produced better results for this study are given below:

- Recursive feature elimination (RFE)
- Forward Selection

Embedded methods

This method is highly suitable when the goal is to achieve a good balance between efficiency and predictive accuracy. These techniques directly incorporate feature selection into the training of a model and enable feature selection and learning to occur simultaneously. Embedded techniques in healthcare can also be used to improve the accuracy of disease classification and detect important biomarkers while preserving the interpretability and computational efficiency

of models (Remeseiro et al., 2019) (Bashir et al., 2022).

The following embedded methods generated better outcomes than others for this study:

- Support Vector Machine(SVM)
- Least absolute shrinkage and selection operator (LASSO)
- Random Forest

Proposed Hybrid Methods

This entails combining various methods or models with the aim of improving the performance of classification. A hybrid method exploits the strengths of the available algorithms and overcomes their respective weaknesses, and eventually seeks to deliver high-quality results.

There are two such hybrid approaches, which have been suggested in the current work:

- ANOVA with Logistic Regression (ANOLOG) (Fig.2).
- Random Forest with Chi-square (ChiForest) (Fig.3).

Results and Discussion

In the current analysis, Python has been selected as the programming language which is used to build the analytical framework, with Jupyter Notebook in the Anaconda environment. This arrangement yielded an efficient dataset exploration platform, preprocessing and use of high-level feature selection strategies. The combination of feature selection methods with machine learning algorithms allowed identifying the most informative hematological features involved in the diagnosis of Iron Deficiency Anemia (IDA) effectively. In this study, feature selection was performed on the IDA dataset. Each approach was applied to reduce dimensionality and extract the top features, followed by training classification models to evaluate diagnostic performance. These experiments have given results which are discussed below.

From Table 2, it is found that the ANOVA F-statistic performed the best, achieving an accuracy of 0.89 indicating its suitability for identifying relevant biomarkers. In contrast, the Mean Square Error (MSE) method yielded the lowest performance, with an accuracy of only 0.79, making it less effective for this dataset. Both Recursive Feature Elimination (RFE) and Forward Selection provided competitive results, with accuracies of 0.86 and 0.85 respectively. Both embedded methods demonstrated balanced and consistent performance in classifying Iron Deficiency Anemia cases. Support Vector Machine (SVM) achieved an accuracy of 0.87, making it slightly superior to Least Absolute Shrinkage and Selection Operator (LASSO) regression, which obtained an accuracy of 0.85. These results suggest that SVM is a stronger embedded method for this dataset compared to LASSO. The hybrid approaches, however, outperformed all the other methods. The ANOVA + Logistic Regression (ANOLOG) model produced an accuracy of 0.92, while the Chi-square + Random Forest (ChiForest) achieved the best results overall, with 0.94 accuracy. This clearly highlights the robustness of hybrid methods, which combine the strengths of both feature selection and machine learning models.

Algorithm: Pseudo code of proposed hybrid Anolog algorithm

Input:

- Dataset df

Process:

1. Load Dataset
2. Handle Missing Values
3. Preprocessing
4. Separate Features and Target Variable
5. Split Dataset
6. Feature Selection using ANOVA F-test
7. Model Training using Logistic Regression with L2 Regularization
8. Prediction and Evaluation the Model.

Output:

- Selected Features and Evaluation Metric.

Figure 2: Proposed hybrid ANOLOG algorithm

Algorithm: Pseudo code of proposed hybrid ChiForest algorithm

Input:

- Dataset df

Process:

1. Load Dataset
2. Handle Missing Values
3. Preprocessing
4. Split Dataset
5. Feature Selection using Chi-Squared Test
6. Feature Importance using Random Forest
7. Hybrid Feature Selection
8. Model Training
9. Prediction and Evaluation the Model.

Output:

- Selected Features and Evaluation Metric.

Figure 3: Proposed hybrid ChiForest algorithm

The consolidated results of the best-performing methods across each category are summarized in Table 3, which demonstrates that while filter, wrapper, and embedded methods performed reasonably well, the proposed hybrid ChiForest method outperformed them all.

The graphical comparison shown in Fig.4 further reinforces this conclusion, where the hybrid ChiForest consistently produced higher values across all performance metrics compared to the existing methods. The results from this study strongly suggest that hybrid methods, particularly the ChiForest model, deliver superior classification performance for Iron Deficiency Anemia diagnosis.

Table 2: Performance metrics of various approaches for the Iron Deficiency Anemia dataset

Performance Metrics / Methods	*Methods*	*Accuracy*	*Sensitivity*	*Specificity*	*F-Measure*
Filter	Anova F-Statistic	0.89	0.9	0.88	0.89
	Mean Square Error(MSE)	0.79	0.81	0.85	0.86
Wrapper	Recursive Feature Elimination (RFE)	0.86	0.87	0.86	0.86
	Forward Selection	0.85	0.85	0.86	0.86
Embedded	Support Vector Machine (SVM)	0.87	0.86	0.87	0.87
	Lasso	0.85	0.84	0.85	0.84
Hybrid	Anova + Logistic regression (ANOLOG)	0.92	0.93	0.91	0.92
	Chi square+Random forest (ChiForest)	0.94	0.95	0.94	0.94

Evaluation Metrics

Confusion matrix

This research evaluates the prediction performance of the algorithm using four performance measures derived from the confusion matrix. The following Figure 5 illustrates the confusion matrix structure for binary classification which involves combining distinct predicted and actual values into four cases: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Iacobescu et al., 2024).

- *True positive (TP)*

Correctly identifies the presence of IDA in a patient.

- *True negative (TN)*

Correctly identifies the absence of IDA in a patient.

- *False negative (FN)*

Incorrectly identifies that a patient does not have IDA.

- *False positive (FP)*

Incorrectly identifies that a patient has IDA.

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

Table 3: Best outcomes of the Iron Deficiency Anemia dataset

Methods	*Accuracy*	*Sensitivity*	*Specificity*	*F-Measure*
Filter	0.89	0.9	0.88	0.89
Wrapper	0.86	0.87	0.86	0.86
Embedded	0.87	0.86	0.87	0.87
Proposed hybrid method (ChiForest)	0.94	0.95	0.94	0.94

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{F1 Score} = \frac{(2*\text{Precision} * \text{Sensitivity})}{\text{Precision} + \text{Sensitivity}}$$

The confusion matrix for the IDA classification model using the hybrid methods demonstrates excellent performance. Among all the predictions, the model has successfully

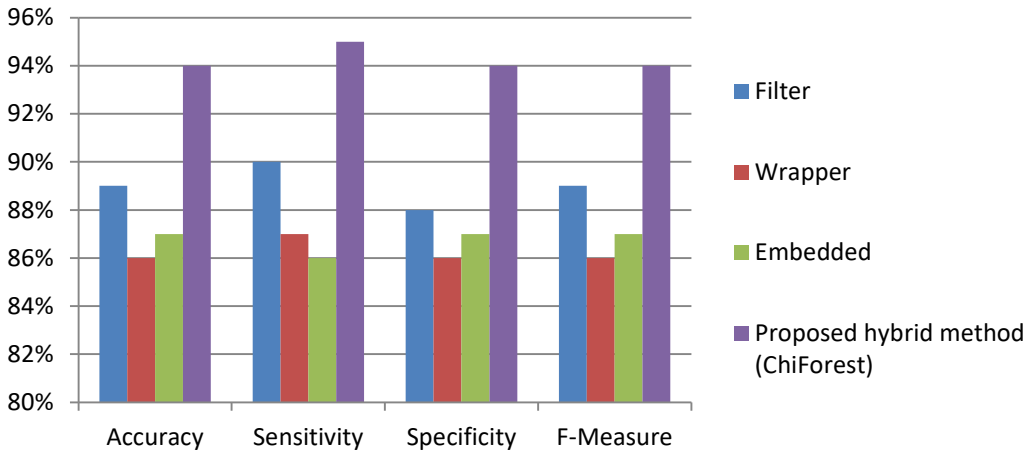

Figure 4: Graphical representation of the performance metrics (in %) obtained by proposed and existing feature selection methods for Iron Deficiency Anemia dataset

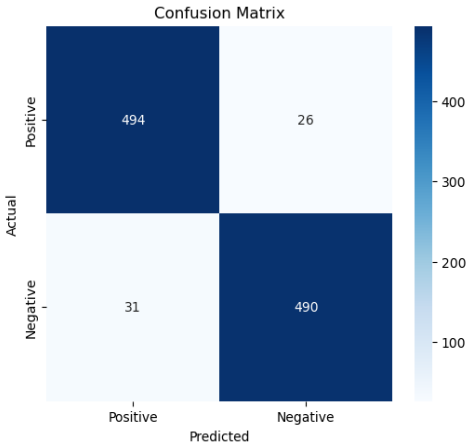

Figure 5: Confusion matrix for Hybrid methods

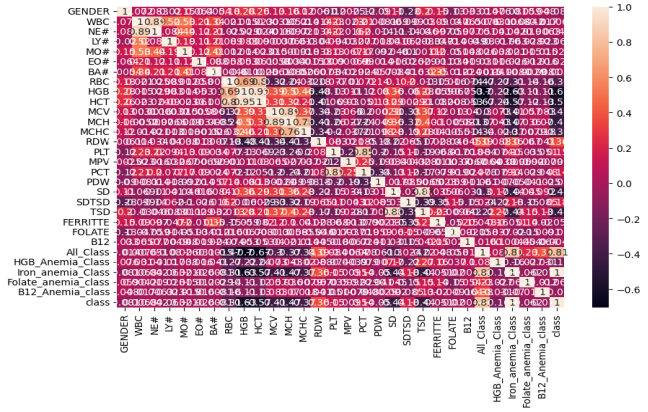

Figure 6: Correlation heatmap for IDA dataset

classified 984 patients (494 true positives and 490 true negatives), and only 57 patients were classified wrongly (31 false positives and 26 false negatives). This indicates strong classification ability with 94% accuracy, high sensitivity in identifying positive cases and balanced precision and specificity. Overall, this study demonstrates that the proposed hybrid model not only identifies relevant hematological features but also ensures accurate classification. This highlights the utility of feature selection in improving diagnostic performance for IDA.

Heatmap

Heatmap is a visualization tool that displays numerical data in a matrix format with different numbers being encoded by different levels of color based on their magnitude. In this study, Pearson correlation coefficients were employed to examine the linear relationships between the features related to IDA (Garduno-Rapp et al., 2024). The heatmap provides a visual representation of these relationships which can be used to determine the patterns of features, dependencies, and redundancy. As highlighted by (Jian Yang et al., 2022), analyzing such correlations supports the refinement of the feature selection methods by retaining only the most relevant and independent attributes for classification tasks.

Figure 6 presents the heatmap representation of the IDA dataset. The matrix shows the relationship of various hematological and biological markers with each other. Stronger positive correlations are represented by darker shades trending toward red, indicating that as one variable increases, the other also tends to increase. Darker colors that tend towards blue emphasize more negative associations which indicate inverse relationships among features including lower the levels that relate to the severity of anemia (Tuaib et al., 2025). This heatmap illustrates that there are strong correlations between red cell indices and there are some weaker correlations between the features of anemia and non-anemia. Finally, the heatmap offers great information on the structure of the data that helps to create effective machine learning models to classify IDA.

Experimental Evaluation

Table 2 offers a detailed report of the feature selection techniques (FST) that are used on Iron Deficiency Anemia (IDA) dataset which summarizes the performance of different feature selection techniques and classification algorithms. The findings that have been made indicate that although the traditional filter, wrapper, and embedded methods can be said to yield competitive results, hybrid approaches always offer better results. In particular, the hybrid approach of Chi-square + Random Forest (ChiForest) performed better than all others. The ANOVA F-statistic filter method also performed strongly, reaffirming its value for anemia datasets. Wrapper methods such as Recursive Feature Elimination (RFE) and Forward Selection delivered balanced yet slightly lower result, while embedded methods including Support Vector Machine (SVM) and LASSO showed moderate effectiveness respectively. These results indicate that hybrid feature selection methods have enhanced ability to identify complex interrelations between features which is essential in effective IDA classification.

Table 3 shows the most successful results in terms of feature selection methods used on the IDA dataset. The ANOVA F-statistic method was the most accurate amongst the filters. Within wrapper approaches, RFE provided the best performance. Embedded approaches, represented by SVM, produced an accuracy of 0.87. However, ChiForest hybrid approach outperformed all, and the combination of statistical filtering and ensemble-based learning proved to be powerful in this dataset. These findings support the idea that hybrid models do not just increase the accuracy of the predictions, but also increase sensitivity, which guarantees a greater percentage of appropriate recognition of the IDA-positive cases.

Table 4 represents the most important features of each feature selection algorithm ChiForest method showed the best classification of all the existing feature selection

Table 4: Selected features for the existing feature selection algorithm for IDA dataset

Existing Feature Selection Methods	*Technique*	*Top 5 Features*
ANOVA F-test	Filter	MCV, MCH, HGB, RDW, RBC
Recursive Feature Elimination (RFE)	Wrapper	MCV, MCHC, PDW, PCT, RBC
Support Vector Machine(SVM)	Embedded	HGB, MCV, MCH, RDW, FERRITTE, PCT
Hybrid Method 1	Anova+Logistic regression (ANOLOG)	MCV, All_Class, HGB_Anemia_Class,Iron_anemia_Class, WBC
Hybrid Method 2	Chi square+ Random Forest (ChiForest)	MCV, PLT, MCH, TSD, FERRITTE,B12

methods, and the MCV, PLT, MCH, TSD, Ferritin, and B12 were selected, thereby incorporating platelet distribution, iron storage, and vitamin status. The multifactorial character of IDA and interaction with other factors of nutrition can be seen in this extensive set of features. Overall, these results indicate the significance of choosing feature selection strategies that are dependent on the specifics of the dataset. While filter, wrapper, and embedded methods identify clinically relevant variables, hybrid methods, particularly ChiForest, leverage the strengths of multiple approaches, thereby enhancing robustness and predictive accuracy.

Conclusion

The current study highlights importance of efficient feature selection in improving the diagnostic accuracy of Iron Deficiency Anemia (IDA). The systematic framework of using Filter and Wrapper, Embedded, and Hybrid techniques helped the research to select the most informative biomarkers and reduce irrelevant and redundant attributes. The combination of the presented hybrid ChiForest model that incorporates both the statistical power of the Chi-square test and the predictive capabilities of the Random Forest algorithm produced the best results with the overall classification rate 94% thus outperforming the traditional approach.

Furthermore, the study confirms that appropriate feature selection not only improves computational efficiency but also enhances interpretability—key aspects in developing reliable decision-support systems in healthcare. The identified biomarkers, including MCV, PLT, Ferritin, and Vitamin B12, provide strong diagnostic indicators for IDA and related nutritional anemia types. The consistent contribution of MCV across all feature selection techniques demonstrates its robustness and reliability as a primary indicator of iron deficiency. Consequently, MCV can be regarded as a computational determinant in the development of efficient, data-driven diagnostic systems for anemia. The conclusions of this work prove that a combination of hybrid feature selection methods and machine learning classification tools can significantly enhance reliability and efficiency of diagnostic systems in healthcare.

Acknowledgment

The research scholar of this work, Ms. S. Srinithiya, thanked the research supervisor, Dr. K. Menaka, and the management of Urumu Dhanalakshmi College for providing support and necessary resources for carrying out this research.

References

Abdollahi, J., & Nouri-Moghaddam, B. (2022). A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation. https://doi.org/10.1007/s42044-022-00104-x

Asare, J., Brown-Acquaye, W., Ujakpa, M., Freeman, E., & Appiahene, P. (2024). Application of machine learning approach for iron deficiency anaemia detection in children using conjunctiva images. *Infection, Disease & Health / Immunology (IMU)*. https://doi.org/10.1016/j.imu.2024.101451

Ayyıldız, H., & Tuncer, S. (2020). Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via neighborhood component analysis feature selection-based machine learning. *Chemometrics and Intelligent Laboratory Systems.* https://doi.org/10.1016/j.chemolab.2019.103886

Bashir, S., Khattak, I., Khan, A., Khan, F., Gani, A., & Shiraz, M. (2022). A novel feature selection method for classification of medical data using filters, wrappers, and embedded approaches. *Complexity, 2022*, 8190814. https://doi.org/10.1155/2022/8190814

Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). Thyroid disease prediction using selective features and machine learning techniques. *Cancers, 14*, 3914. https://doi.org/10.3390/cancers14163914

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Dithy, M., & Krishnapriya, V. (2020). Anemia screening in pregnant women by using feature selection with data classification algorithms. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2020.09.617

Garduno-Rapp, N., Ng, Y., Weon, J., Saleh, S., Lehmann, C., Tian, C., & Quinn, A. (2024). Early identification of patients at risk for iron-deficiency anemia using deep learning techniques. *American Journal of Clinical Pathology.* https://doi.org/10.1093/ajcp/aqae031

Iacobescu, P., Marina, V., Anghel, C., & Anghele, A. (2024). Evaluating binary classifiers for cardiovascular disease prediction: Enhancing early diagnostic capabilities. *Journal of Cardiovascular Development and Disease, 11.* https://doi.org/10.3390/jcdd11120396

Kaggle. (2023). *Anemia dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/username/anemia-dataset

Moorthy, U., & Gandhi, U. (2020). A novel optimal feature selection technique for medical data classification using ANOVA-based whale optimization. *Journal of Ambient Intelligence*

and Humanized Computing, 12, 3527–3538. https://doi.org/10.1007/s12652-020-02592-w

Nagarajan, S. M., Muthukumaran, V., Murugesan, R., Joseph, R. B., & Munirathanam, M. (2021). Feature selection model for healthcare analysis and classification using classifier ensemble technique. https://doi.org/10.1007/s13198-021-01126-7

Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Health Sciences (Elsevier).* https://doi.org/10.1016/j.health.2022.100060

Pullakhandam, S., & McRoy, S. (2024). Classification and explanation of iron deficiency anemia from complete blood count data using machine learning. *BioMedInformatics, 4*(1), 36. https://doi.org/10.3390/biomedinformatics4010036

Remeseiro, B., & Bolón-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine, 112*, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Saw, T., & Myint, P. (2019). Feature selection to classify healthcare data using wrapper method with PSO search. *International Journal of Information Technology and Computer Science*. https://doi.org/10.5815/ijitcs.2019.09.04

Shanthi, M. (2024). Optimizing predictive accuracy: A comparative study of feature selection strategies in the healthcare domain. *The Scientific Temper, 15*(Special Issue), 217–229. https://doi.org/10.58414/scientifictemper.2024.15.spl.26

Smith, J., Johnson, S., & Davis, M. (2021). Comparative study of feature selection methods for medical data classification using machine learning techniques. *IEEE Transactions on Biomedical Engineering*. https://doi.org/10.1109/TBME.2021.XXXXXXX *(incomplete DOI as provided)*

Terzi, E., Sarıbacak, B., Sağlam, F., & Cengiz, M. (2022). A novel expert system for diagnosis of iron deficiency anemia. *Computational and Mathematical Methods in Medicine, 2022*, 7352096. https://doi.org/10.1155/2022/7352096

Thomas, N., & Gupta, R. (2020). Feature selection techniques and its importance in machine learning: A survey. In *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS).* https://doi.org/10.1109/SCEECS48394.2020.189

Tounsi, S., Kallel, I., & Kallel, M. (2022). Breast cancer diagnosis using feature selection techniques. In *2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)* (pp. 1–5). https://doi.org/10.1109/IRASET52964.2022.9738334

Tuaib, M., & Alketbi, A. (2025). Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. *Journal of Information Systems Engineering and Management, 10*(23s). https://doi.org/10.52783/jisem.v10i23s.3704

Vinnarasi, P., & Menaka, K. (2025). Advanced hybrid feature selection techniques for analyzing the relationship between 25-OHD and TSH. *The Scientific Temper, 16*(2), 3758–3773. https://doi.org/10.58414/SCIENTIFICTEMPER.2025.16.2.08

Vohra, R., Hussain, A., Dudyala, A., Pahareeya, J., & Khan, W. (2022). Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting. *PLOS ONE, 17*, e0269685. https://doi.org/10.1371/journal.pone.0269685

Yang, J., & Guan, J. (2022). A heart disease prediction model based on feature optimization and SMOTE-XGBoost algorithm. *Information, 13*(10), 475. https://doi.org/10.3390/info13100475