Data Science and Engineering
https://doi.org/10.1007/s41019-024-00257-8

RESEARCH PAPER

An Optimal Model for Medical Text Classification Based on Adaptive Genetic Algorithm

Ghada Ben Abdennour[1] · **Karim Gasmi**[2,3] · **Ridha Ejbali**[1]

Received: 5 February 2024 / Revised: 6 June 2024 / Accepted: 14 July 2024

Abstract
Automatic text classification, in which textual data is categorized into specified categories based on its content, is a classic issue in the science of Natural Language Processing. In recent years, there has been a notable surge in research on medical text classification due to the increasing availability of medical data like patient medical records and medical literature. Machine learning and statistical methods, such as those used in medical text classification, have proven to be highly efficient for these tasks. However, a significant amount of manual labor is still required to categorize the extensive dataset utilized for training. Recent research have demonstrated the effectiveness of pretrained language models, including machine learning models, in reducing the time and effort required for feature engineering by medical experts. However, there is no statistically significant enhancement in performance when directly applying the machine learning model to the classification task. In this paper, we present a hybrid machine learning model that combines individual traditional algorithms augmented by a genetic algorithm. However, the improved model is designed to enhance performance by optimizing the weight parameter. In this context, the best single model demonstrated commendable accuracy. In addition, when applying the hybridization approach and optimizing the weight parameters, the results were substantially enhanced. The results underscore the superiority of our augmented hybrid model over individual traditional algorithms. We conduct experiments using two distinct types of datasets: one comprising medical records, such as the Heart Failure Clinical Record and another consisting of medical literature, such as PubMed 20k RCT. So, the objective is to clearly showcase the effectiveness of our approach by highlighting the significant enhancements in accuracy, precision, F1-score and Recall achieved through our improved model.

Keywords Medical text classification · Ensemble learning · Optimization · Genetic algorithm

1 Introduction

Artificial intelligence (AI), especially machine learning (ML), plays a crucial role in text classification. Text classification is a fundamental task in Natural Language Processing (NLP) widely utilized in various real-world applications such as spam detection [25], language identification [55] and sentiment analysis [43]. Each of these applications deals with a different type of document and class, highlighting the versatility and significance of text classification in diverse contexts.

In contemporary healthcare, the integration of complex, extensive and diverse health data is indispensable for modern medicine. This includes various sources like patient medical records, e-health, results from radiological imaging, telemedicine and medical literature. The emphasis of this paper is on a particular category of health data, namely, textual data. So, it is important to note that the analysis

Karim Gasmi and Ridha Ejbali have equally contributed to this work.

✉ Ghada Ben Abdennour
ghadabenabdennour@gmail.com

Karim Gasmi
kgasmi@ju.edu.sa

Ridha Ejbali
ridha_ejbali@ieee.org

[1] Research Team in Intelligent Machines (RTIM), National School of Engineering of Gabes, Gabes University, Gabes, Tunisia

[2] Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakkaka, Saudi Arabia

[3] Research laboratory on Development and Control of Distributed Applications (REDCAD), ENIS, Sfax University, Sfax, Tunisia

Published online: 31 July 2024

of medical texts has become a crucial area of research for health professionals and researchers. ML has been proven to be immensely helpful in addressing healthcare-related challenges due to its strength and success in recent times through the techniques of NLP.

Medical text contains medical records and medical literature; the former is a record of the medical activity process of the doctor's examination, diagnosis, treatment and development of the patient's disease. For example, cardiovascular diseases (CVDs) encompass a spectrum of heart and blood vessel disorders, including coronary heart disease, cerebrovascular diseases and heart failure (HF), contributing to approximately 17 million global deaths annually, accounting for 31% of global mortality (World Health Organization) [2, 13]. So, heart failure (HF) manifests when the heart fails to adequately pump blood and electronic medical records of patients quantify symptoms, body metrics, clinical laboratory test values and associated conditions such as diabetes, high blood pressure, HIV, alcohol abuse, thyroid disorders, radiation, chemotherapy, etc. So, it describes the patient's medical history; it is detailed information about the patient during treatment.

Conversely, medical literature represents a compendium of research outcomes, elucidating the latest medical methodologies in scientific articles. In the last ten years, over 50 million academic publications have been released and the yearly output of articles continues to increase [29, 35, 37]. The National Library of Medicine in the United States manages MEDLINE, which indexes roughly half of these medical publications [39]. Thus, the healthcare literature is very wide, with millions of scientific articles found on websites like PubMed [15], Google scholar [1] ...etc.

In brief, with the intention of demonstrating the generality of the approach presented in this paper, we chose two datasets of different types, as mentioned and discussed at the top.

Given this burgeoning medical data, the accurate classification and analysis of medical texts have become imperative for advancing medical research. Consequently, this study aims to develop an NLP-based text classification system to automatically assign PubMed abstracts and to predict the survival of patients with heart failure in the follow-up period.

Furthermore, it's worth mentioning that the important clinical information resource contains complex medical vocabularies, so the classification in the medical domain is more challenging than those in other domains. So, it is difficult to find a classification model that performs well in both medical records and medical literature. To address these challenges, this paper proposes a hybrid machine learning model augmented by a genetic algorithm (GA) to automatically optimize weight parameters with the aim of enhancing performance and the experimental validations are performed on two types of medical text datasets: medical records and medical literature datasets, affirming the efficacy of our approach. This study includes several stages, starting with data pre-processing and feature extraction, then moving on to a thorough comparative analysis using five classifiers (Support Vector Machine, Random Forest, Decision Tree, Logistic Regression and Naïve Bayes). A combination of three top-performing algorithms is paired together using different ensemble learning techniques and a specific selection of weight parameters based on an optimization algorithm is automatically chosen. Further, the validation through a comparative study against state-of-the-art text classification methods on these two datasets validates the effectiveness of our proposed method.

Ultimately, the main contribution of this study is the improvement of medical text classification models using a hybrid approach. This approach combines the strengths of different traditional machine learning models within a soft voting classifier. The proposed method automatically optimizes the weight parameters of the model. This automation surpasses the manual optimization techniques used in previous studies, such as those by Ben Abdennour et al. [7], by allowing for precise and optimal weight values based on each model's contribution. Consequently, the benefits of each individual model are fully used.

The remainder of the paper is organized as follows: Sect. 2 introduces a literature review on text classification in both general and medical domains in the area of machine learning. The proposed method is described in detail in Sect. 3. Section 4 encompasses an evaluation and analysis of the proposed method via experiments conducted on two different types of datasets. These experiments encompass details on the hybrid strategy, the weight parameter selection and the resulting experimental outcomes. Finally, in Sect. 5, we conclude the paper.

2 Related Work

In this context, we will explore the latest strides and practical applications of ML in both domain general and medical text classification, uncovering their methodologies and applications.

2.1 General Text Classification

General text classification with machine learning involves the application of various algorithms to categorize text data into predefined classes. Techniques such as NLP and supervised learning models have been extensively utilized. These methods analyze textual content by cleaning,

extracting features and learning from labeled data to classify documents, emails, articles or social media posts into pertinent classes based on their content. So, we aim to provide an overview of the research efforts.

Firstly, Rustam et al. [47] propose a machine learning-based approach for spam email detection with good performance. Preprocessing steps include punctuation, number and stop word removal, along with conversion to lowercase, stemming and lemmatization. Combining bag of words (BoW) and term frequency-inverse document frequency (TF-IDF) is a proposed method for feature fusion. Additionally, to reduce the impact of data imbalance on models' overfitting, random under-sampling is used on the majority class. Subsequently, multiple machine learning models are utilized for classification purposes, namely random forest, gradient boosting machine, support vector machines, Gaussian Naïve Bayes and logistic regression. The study's findings highlight random forest and logistic regression as the most effective models among the ones investigated.

Besides, the debate and controversy surrounding the COVID-19 vaccine have influenced many people's decisions to either accept or refuse vaccination. Qorib et *al.* [44] explore public sentiments regarding COVID-19 vaccine hesitancy. The study looks at five machine learning algorithms: Random Forest, Logistic Regression, Decision Tree, LinearSVC and Naïve Bayes. It does this by using different combinations of text analysis methods on an English Twitter dataset about the COVID-19 vaccine. The results indicate that combining TextBlob, TF-IDF and LinearSVC achieves the highest accuracy of 96.75%.

Furthermore, Naeem et al. [41] present a method to analyze sentiments within movie reviews using supervised machine learning classifiers. The aim is to help individuals choose movies based on the reviews' popularity and interest. Four machine learning algorithms (Decision Tree, Random Forest, GBC and Support Vector Machines) are applied for sentiment analysis, trained on preprocessed datasets. Additionally, four feature extraction methods, such as BoW, TF-IDF, GloVe and Word2Vec, are explored to identify impactful and meaningful review features.

Additionally, another research endeavor by Luo et al. [38] applied a ML model for the classification of English texts and documents. The study conducted a comparative analysis of various machine learning algorithms, including Naïve Bayes, Support Vector Machine and Logistic Regression. In the initial phase, they executed preprocessing to identify optimal features, such as frequency, initial letter, paragraph, question markand full stop. Subsequently, they extracted text features from documents. The simulation results unequivocally demonstrated that Support Vector Machine surpassed the performance of the other machine learning models.

2.2 Medical Text Classification

In light of the diverse applications and methodologies in general and medical text classification, it's clear that machine learning is still making a big impact and advancing our understanding and capabilities. Transitioning from general text classification to focusing specifically on medical text classification, we can observe how machine learning models advance healthcare and become increasingly apparent in various tasks within the health field. So, with an increasing volume of medical literature, patient records and clinical notes, machine learning models tailored for medical text analysis play a pivotal role in tasks such as disease diagnosis, prognosis prediction, etc.

For example, the researchers Chadaga et al. [10] employed blood tests in conjunction with a ML model to predict COVID-19 infection, achieving a commendable classification accuracy of 91%. To enhance the quality of the data, a preprocessing step was implemented for data cleaning. The classification task involved the use of four distinct classifiers: XGBoost, Random Forest, K-Nearest Neighbors and Logistic Regression. Additionally, a dataset balancing technique known as the Synthetic Minority Oversampling Technique (SMOTE) was applied. Furthermore, the Shapley Additive Explanations (SHAP) method played a pivotal role in assessing the significance of each feature in distinguishing COVID-19 infection within the dataset.

Secondly, Chang et al. [11] propose an e-diagnosis system that leverages ML algorithms for predicting diabetes mellitus. The Pima Indians diabetes dataset is employed to train and test three classifiers: Naïve Bayes, Random Forest and J48 decision trees. The study thoroughly analyzes the performance of each classifier, aiming to identify the one that excels based on various evaluation metrics.

Moreover, Kambar et al. [31] aim to classify drug interactions in medical documents using machine learning techniques. As part of the preprocessing steps, they eliminated stop words, punctuation, excessive white space, unwanted characters and words with no meaningful information. They employed TF-IDF for feature extraction. The study evaluates the performance of several ML models, including Random Forest, Logistic Regression, Support Vector Machine, XGBoost and Decision Tree. Among these models, the Decision Tree model exhibited the most promising results, achieving an accuracy of 95%, a recall of 100%, F1-score of 92% and a precision of 86%. To bridge the research findings in this state-of-the-art, recent studies have showcased the prowess of ML. For instance, in the domain of autism research, a machine learning-based approach was developed by Uddin et al. [53] to effectively detect and identify autism spectrum disorder (ASD) traits. The authors used the SMOTE method, along with feature transformation and selection techniques, to address dataset imbalances.

Subsequently, various classification methods were applied alongside hyperparameter optimization. Ultimately, the AdaBoost method emerged as the most effective classifier, yielding the best outcomes in their analysis.

Previous research has examined the medical texts classification using datasets similar to the ones used in our suggested method, specifically the PubMed and Heart Failure clinical record datasets. For instance, Anantharaman et al. [4] conducted a study where they applied tokenization, stemming and lemmatization techniques, and eliminated stop words as a preprocessing measure. Following the preprocessing procedures, they have introduced a collection of words, TF-IDF, and various topic modeling techniques (LDA, LSA and NMF) as part of the feature extraction process. RF is the classification model. The analysis yielded an accuracy ranging from 48.6 to 57.5% for the PubMed 20K RCT dataset. While a study by Mercadier [40] has talked about abstract classification using the PubMed 200K RCT dataset, a comparative study between diverse models of ML is done with the removal of punctuation, special characters and stop words and the conversion of all the text to lowercase as a preprocessing step. Moreover, TF-IDF is the feature extraction method used. Mercadier achieved different values of accuracy, but the best is equal to 65.20%.

However, Chicco et *al.* [13] employed multiple ML classifiers to predict patient survival and rank the pivotal risk features. Their feature rankings revealed serum creatinine and ejection fraction as the primary attributes crucial for constructing the prediction model. In their comparative analysis, using the entire feature set yielded an accuracy of 74%, whereas focusing solely on two features, serum creatinine and ejection fraction, significantly enhanced accuracy to 83.8%. Afterwards, Firas et al. [32] propose an approach that combines various ML techniques with Shapley values. Their objective is to increase the risk coefficients applied by Shapley with the k-fold technique in order to maximize the reliability of the explainability. Their findings highlight the significance of ejection fraction and serum creatinine as pivotal features for predicting patients at risk of mortality during hospital follow-up. This means that the k-fold method with Shapley values made it easier to rank the importance of features and achieved an accuracy of 83.3% in predicting patient survival.

Alternatively, the KNN algorithm stands out as one of the most extensively employed ML techniques. Uddin et al. [54] look at different types of KNN in their study, such as the Classic, Adaptive, Locally Adaptive, k-Means Clustering, Fuzzy, Weight-Adjusted, Mutual, Ensemble, Hassanat and Generalized Mean Distance models. Their investigation revolves around comparing the predictive performance of these variants in forecasting heart failure disease. Ensemble approach KNN can be selected as the most suitable KNN variants according to their high accuracy, precision and recall metrics.

3 Methodology

In this section, we will discuss our proposed model for medical text classification that we have developed. Figure 1 depicts the workflow of our approach. At the beginning, the datasets containing medical text from two distinct sources are loaded for classification. This text is then preprocessed to enhance its quality. Before generating the input for the classifier, we use different techniques for text representation. In the classification phase, we conduct a comparative analysis among different models, utilizing two strategies (single and hybrid). Our objective here is to determine the most effective classifier based on Accuracy, Precision, F1-score and Recall metrics. Finally, we augment the best-performing classifier with a selection of weight hyperparameters based on a genetic algorithm, aiming to optimize the model further. Each of these steps will be elaborated upon extensively in the subsequent sections.

3.1 Pre-processing

At the forefront of text classification in NLP, pre-processing stands out as an important and major step. It is used to transform the data into a useful and efficient format so that it can be fed into the ML model used and reduce the error during the classification.

For Pubmed 20K RCT dataset, this process involves tokenization, which involves strings that are divided into smaller tokens, or, in our case, tokenizing sentences into words. After that, we convert all the text into lowercase letters in order to unify all the text so that the classification becomes easier. In addition, there are some words that have no context in the text; these words are known as stop words, so it necessitates the removal of stop words like and, is, of, on, a, etc. Furthermore, remove punctuation from the text because it is not useful in our case.

On the other hand, for the Heart Failure clinical record dataset, the number of deaths (death event = 0) and survival (death event = 1) are 96 and 203, respectively (out of 299 patients). In statistical terms, there are 32.11% positives and 67.89% negatives. So, it produces class imbalanced data that can't be used with ML algorithms to achieve the desired performance. To make our predictions more realistic and solve the problem of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) [9, 18, 27, 42] has been utilized to balance the class distributions. It has emerged to be an effective method when there is an imbalance in the distribution of classes.

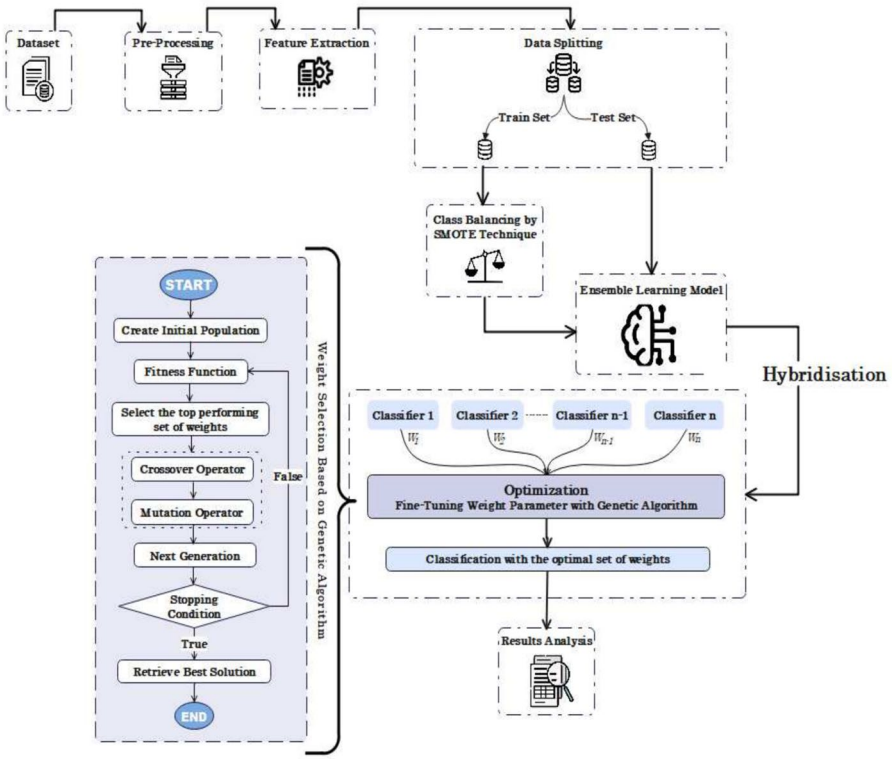

Fig. 1 Proposed model for medical text classification

3.2 Feature Extraction

Once the data has been pre-processed and cleaned, the crucial step in text classification is extracting the relevant features. In this stage, the text is transformed into numerical features that may be readily comprehended and processed by ML models. The feature extraction employed in this study for the literature dataset include CountVectorizer (CV) and Term Frequency-Inverse Document Frequency (TF-IDF).

- *Term Frequency - Inverse Document Frequency (TF-IDF)* The objective of this work is to assess the significance of words in textual documents or a collection of texts. Term frequency (TF) quantifies the frequency of a word in a document by dividing the number of occurrences of the word by the total number of words in the document. The IDF is calculated by taking the logarithm of the total number of documents in the corpus and dividing it by the number of documents that include the specific word [6, 8, 30, 52, 58]. In Eq. (1), for a document or corpus doc, the vector V is computed as:

$$V = \mathrm{TF}(t, \mathrm{doc}) \times \mathrm{IDF}(t) \tag{1}$$

Where $\mathrm{TF}(t, \mathrm{doc})$ represents the frequency of term t appearing in document docand $\mathrm{IDF}(t)$ can be calculated based on word frequencies across the corpus, according to Eq. (2):

$$\mathrm{IDF}(t) = \log\left(\frac{|D|}{\mathrm{DF}(t)}\right) \tag{2}$$

The term $|D|$ stands for the total number of documentsand $\mathrm{DF}(t)$ denotes the number of documents where term t appears at least once.

- *CountVectorizer (CV)* The text document is transformed into a matrix, with each row representing a document and each column representing a word (token). The value within the matrix indicates the frequency of each word in each document. This method transforms the textual document into numerical characteristics, which can be utilized as input for ML algorithms [6, 52, 57].

3.3 Selection Model

This research utilizes various ML classifiers in two distinct procedures (single and hybrid) and compares them to determine the optimal strategy and classifier.

3.3.1 Single Model

(1) *Support Vector Machine (SVM)* SVM is a supervised learning algorithm extensively employed for classification tasks, notably recognized for its efficacy in text classification. The primary objective of SVM is to discern a hyperplane that effectively segregates data into two or more classes. This hyperplane is strategi-

cally positioned to maximize the geometric margin, the distance between the hyperplane and the nearest training data points from each class. SVM is a robust algorithm capable of handling both linear and non-linear data by transforming it into a higher-dimensional space. By judiciously selecting the hyperplane, SVM ensures optimal classification. However, Eq. (3) symbolizes the fundamental mathematical representation governing this classification process. It delineates the decision-making mechanism of SVM in establishing a boundary between classes, ensuring an effective separation of data points [3, 5, 17, 26, 30].

$$f(x) = w^T x + b \tag{3}$$

Where:

- w = dimensional coefficient;
- b = offset.

(2) *Naïve Bayes (NB)* NB is a commonly used model for text classification problems that is known for its optimality and efficiency. It is a probabilistic classifier that refers to the Bayes Theorem. See Eq. (4).

$$P(y \mid x_1, ..., x_j) = \frac{P(x_1, ..., x_j \mid y) \cdot P(y)}{P(x_1, ..., x_j)} \tag{4}$$

Where:

- P(x1,...,x j | y) = Likelihood;
- P(x1,...,xj) = Normalization constant;
- P(y) = Prior.

The Bayes' theorem establishes a relationship between a given class variable y and a dependent feature vector x_1 through x_j. Initially, NB calculates the prior probability of each category based on the training data, which is the probability of each category occurring in the dataset without considering any of the input features. Then, when we presented a new input data point, it calculated the conditional probability of each category given the input features and assigned the predicted category that got the highest probability. It used the statistics on each class to evaluate the weights of the model [12, 19, 30, 34, 46, 56].

(3) *Decision Tree (DT)* A DT is a supervised ML method that is commonly utilized for classification. In tree structure, there are three components: nodes (root or internal), branches and leaf nodes. The internal nodes correspond to test data, that is, the attributes to classify. The branches represent the result of the test. The leaf nodes signify the final decision (class). The idea is that it iteratively selects the best attribute, the decision node and that the classification process starts at the root of the tree and proceeds down the tree according to the answers to the tests that label the internal nodes. The resulting class is determined by the majority class associated with the leaf node that corresponds to the input description [5, 20, 45]. In the context of DT algorithms, especially in classification tasks, entropy serves as a pivotal criterion for determining how to effectively split data. This measure plays a crucial role in deciding the optimal attribute for data partitioning. So, entropy, in general, is given by Eq. (5).

$$\text{Entropy} = -\sum_{i=1}^{N} p_i \cdot \log_2(p_i) \tag{5}$$

Where:

- N = number of classes;
- Pi = proportion of samples belonging to class i.

(4) *Random Forest (RF)* The RF classifier is a popular ML method that is highly effective for solving text classification tasks. By using random subsets of the data, a multiple-decision tree constructs it. In the prediction phase, the algorithm processes each data point through all the decision trees and combines their results to generate the final prediction. Typically, the final prediction is determined by the majority vote of all the decision trees [5, 34, 49].

(5) *Logistic Regression (LR)* This algorithm has gained importance in recent times due to its multiple advantages. LRis a family of regression analyses used to model the probability of a certain outcome based on one or more predictor variables [22, 34, 43, 49]. This statistical technique is primarily used for classification problems, where the output variable is a binary variable that takes on one of two possible values. It employs the sigmoid function to map each data point. Eq. (6) illustrates the sigmoid function.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

3.3.2 Ensemble Learning Model

Ensemble learning involves training multiple individual classifiers (learners) to predict a solution for the same problem. Some hybrid techniques create homogeneous learners using a single learning algorithm, while others form heterogeneous base learners by employing different learning algorithms. Typically, the latter is particularly beneficial when leveraging the advantages and robustness of each learning algorithm. So, to attain a high performance in classification

in ensemble learning, learners must be accurate and diverse. To clarify, multiple learners are taking an instance vector as input and a technique is employed to aggregate or combine their results to result in a single output [16]. This paragraph focuses on the hybridization strategy in the field of ML, which involves combining several methodologies. Multiple combination approaches are documented in the literature, and we will now present the most commonly employed techniques in this evolving domain, including voting, bagging and stacking.

(1) *Voting* A voting classifier can utilize two voting techniques: hard and soft. During hard voting, each classifier casts a vote for the output class, and the class that receives the majority of votes is selected. Conversely, the soft voting classifier employed in our approach amalgamates the results of various distinct classifiers to generate a conclusive forecast. Each classifier assigns a probability to the output class, and these probabilities are combined and weighted to determine the most likely class, which is then selected as the forecast [5, 33, 34]. (See Fig. 2).

In the scenario depicted in Fig. 2, the contribution of each individual classifier is equal, indicating that the weight parameter W is set uniformly across all classifiers.

Mathematically speaking, the soft voting classifier is denoted by Eq. (7), which is expressed as follows:

$$S = \text{argmax}_i \sum_{j=1}^{n} W_i * p_{ij} \tag{7}$$

Where:

- S: Final prediction;
- argmax function: Return class with the highest probability;
- i: Number of classes, $i = \{1, 2, \ldots, m\}$;
- j: Number of models, $j = \{1, 2, \ldots, n\}$;
- W: Represents the weight;
- P: Probability from the classifiers.

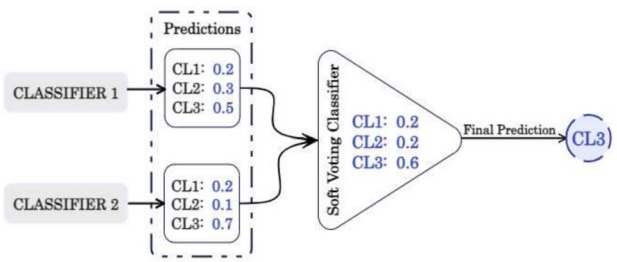

Fig. 2 Soft voting classifier with equal contributions ($W_1 = W_2$)

2. *Bagging* Ensemble machine learning commonly employs a sampling approach called bagging, which involves creating subsets of the original training set. Each model in the ensemble is trained separately on one of these subgroups. These models function concurrently, generating predictions independently. Ultimately, the ensemble amalgamates the forecasts generated by all the models to ascertain the ultimate prediction [51].
3. *Stacking* Stacking is an ensemble machine learning strategy that entails training many models on the identical dataset. The forecasts generated by these models are subsequently used as input characteristics to design a novel matrix. This matrix is employed to train a concluding model, referred to as a meta-learner, which acquires the ability to amalgamate the forecasts from the various models [48].

3.3.3 Optimal Parameter Selection Using Genetic Algorithm (GA)

The Genetic Algorithm (GA) adheres to the principles of natural selection proposed by Darwin's theory. It is an optimization algorithm that draws inspiration from the mechanisms of natural selection and genetics.

The GA is primarily employed for optimizing search problems and modeling various aspects of optimization, including its application in ML. It operates as a random-based optimization technique within the field of applied mathematics. Optimization, in this particular situation, refers to the process of finding the best possible solutions for both issues that have no limitations and problems that have restrictions. The provided problem's underlying mathematics is defined using mathematical functions and variable expressions. Consequently, the obtained optimal solutions play a crucial role in tasks such as parameter estimation and tuning [21, 28, 36, 50].

In the domain of medical text classification, our focus revolves around optimizing the weight parameters in the soft voting classifier (best classifier), which is a crucial step in improving accuracy. Thus, this optimization significantly influences the task at hand.

GA play a decisive role in optimization. In genetic concept, within the chromosomes, there are genes, and the specific values assigned to these genes are known as alleles. To clarify, each chromosome represents the set of weights for the soft voting classifier and each gene represents the weight value assigned to an individual classifier. Refer to Fig. 3 for an illustration where genes form chromosomes, which are made up of DNA, all contained within the cell's nucleus.

To put it simply, the key components in a genetic algorithm come by operating by generating an initial population of potential solutions, meaning that of chromosomes, where each chromosome represents a set of weights for the soft

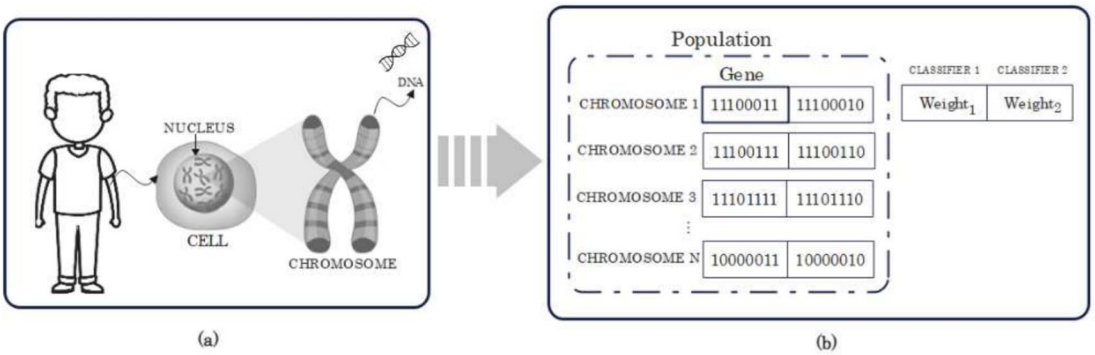

Fig. 3 Correspandence of natural process (**a**) and our strategy (**b**)

voting classifier and these weights determine the contribution of each individual classifier in the ensemble (In our case, each chromosome contains two genes, W1 and W2. Refer to Fig. 3b).

Secondly, evaluate the fitness of each chromosome and of each set of weights in the soft voting classifier, so the fitness function is typically based on performance in terms of accuracy, precision, F1-score and Recall, and the higher-fitness chromosomes are more likely to be chosen, mimicking the concept of natural selection as we said.

Next, apply genetic operators like crossover, which involves merging two sets of weights from two parent solutions to produce offspring; additionally, the mutation operator introduces random changes in the new set of weights for the offspring [36].

In addition, replace the old population with the new population of offspring generated through these two operations. So, these operations shape the succeeding generation by selecting the best-performing candidates, ultimately aiming to achieve an optimal solution across successive generations.

Ultimately, this process continues until a maximum number of generations (stopping criteria) is reached; consult Fig. 1 (left part).

This methodology offers a powerful means to fine-tune the weight parameter, resulting in a model capable of more accurately classifying diverse medical texts, a critical requirement in the healthcare domain.

4 Experiment Results and Discussion

Experiments were carried out to assess the effectiveness of the proposed text classification approach on two distinct dataset types. This part talks about how the experiment was set up, compares the different ML models described in Sect. 3 using two different approaches (single and hybrid) and then talks about the results.

4.1 Dataset Description and Evaluation Metrics

Our model was evaluated on a medical text classification task encompassing both medical record and medical literature classification, utilizing the following datasets:

To initiate this research, we utilized a dataset called PubMed 20k RCT [14]. The dataset was released in 2017 by Dernoncourt and Lee. It is derived from PubMed and is used for classifying sentences in biomedical literature. The collection comprises 20,000 abstracts of randomized controlled trials. Every sentence in each abstract is categorized with one of five labels to indicate its position in the abstract: background, objective, method, result or conclusion. The user's text is simply a backslash character. Furthermore, we utilized a clinical record for Heart Failure [24], which consisted of medical information from 299 individuals diagnosed with heart failure. The aforementioned records were gathered from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad.

Table 1 presents summary statistics for these datasets.

To evaluate the efficacy of our classification technique, we utilized established metrics such as Accuracy, Precision, F1-score and Recall [5, 23].

The equations of the different metrics are described below:

The model's accuracy is a measure of the proportion of correct predictions compared to the total predictions. This can be mathematically represented by Eq. (8).

Table 1 Dataset description

	Type	Size	Classes
PubMed 20K RCT	Literature	20000	5
Heart failure clinical record	Record	299	2

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{8}$$

Precision, another crucial metric, quantifies the proportion of true positive predictions relative to the total number of predicted positive observations. Equation (9) formally defines this metric.

$$Precision = \frac{TP}{(TP + FP)} \tag{9}$$

The recall metric is defined as the ratio of true positive predictions to all observations in the actual class, as shown in Eq. (10).

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

The F1-score, on the other hand, is often considered the harmonic mean of precision and recall and serves as a comprehensive metric for assessing machine learning model performance. Equation (11) outlines its calculation.

$$F1\text{-}score = \frac{2 * Recall * Precision}{Precision + Recall} \tag{11}$$

Where TP, TN, FP and FN mean True Positive, True Negative, False Positive and False Negative, respectively.

4.2 Evaluation of the Feature Extraction Method

Commencing with an assessment of several feature extraction methodologies such as TF-IDF and CV. We employed a distinct classification model for each technique and assessed its performance using diverse evaluation indicators.

The results consistently revealed enhancements in the performance of the classification model with each individual feature extraction technique. Intriguingly, we have noticed that a combination of these technologies can improve the performance of the model overall. It is noteworthy that the extent of improvement varied among different classifiers. This detailed expression suggests that whether combining different feature extraction techniques works well may depend on the complexities of each individual classifier.

Table 2 presents the results of the evaluations of the different feature extraction techniques we discussed.

Therefore, these results highlight that while both TF-IDF and CV individually enhance the performance of classification models, their combination leverages the strengths of both techniques, leading to superior outcomes.

For example, with the TF-IDF feature extraction technique, SVM achieved an accuracy of 77.194%, precision of 76.926%, F1 score of 76.953% and a recall of 77.014%. In contrast, when applying the CV technique, SVM attained an accuracy of 74.478%, precision of 74.652%, F1 score of 74.468% and a recall of 74.469%. On the other hand, when combining the TF-IDF and CV techniques, SVM recorded an accuracy of 78.085%, precision of 77.826%, F1 score of 77.845% and a recall of 77.932%.

This combination is particularly significant for the SVM, RF and LR models, suggesting that the combined approach captures important features from the dataset more effectively.

Table 2 Evaluation feature extraction techniques

	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
SVM				
CV	74.652	74.478	74.469	74.468
TF-IDF	76.926	77.194	77.014	76.953
CV + TF-IDF	**77.826**	**78.085**	**77.932**	**77.845**
DT				
CV	**66.920**	**67.577**	67.466	**67.176**
TF-IDF	65.951	66.516	66.328	66.183
CV + TF-IDF	65.651	66.213	66.198	65.894
RF				
CV	73.932	74.766	73.872	73.514
TF-IDF	74.418	75.281	74.965	74.148
CV + TF-IDF	**74.460**	**75.298**	**75.179**	**74.204**
NB				
CV	**74.534**	**74.308**	**74.288**	**74.276**
TF-IDF	70.019	69.732	69.011	66.264
CV + TF-IDF	69.421	69.029	68.867	65.268
LR				
CV	76.610	77.008	76.902	76.744
TF-IDF	77.284	77.798	77.632	77.406
CV + TF-IDF	**77.284**	**77.798**	**77.522**	**77.406**

The values in bold represent the best results

4.3 Evaluation of the Classification Method

Our main goal was to create an efficient ML model that could accurately classify medical abstracts and make predictions about patient survival over the follow-up period. Consequently, we assessed multiple algorithms utilizing both single and hybrid ML techniques, such as SVM, DT, RF, NB and LR.

Table 3 compares individuals machine learning classification models, evaluating their performance in terms of Accuracy, Precision, F1-score and Recall for the classification task on the two datasets.

Table 3 Comparable result of different algorithms as single strategy; (a) PubMed 20K RCT (b) heart failure clinical record

	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
a				
SVM	77.826	78.085	77.932	77.845
DT	65.651	66.213	65.933	65.894
RF	74.460	75.298	75.179	74.204
NB	69.421	69.029	65.786	65.268
LR	77.284	77.798	77.522	77.406
b				
SVM	77.348	75.555	76.023	76.111
DT	68.520	68.888	68.888	68.565
RF	79.873	78.888	78.697	78.038
NB	61.404	54.444	55.139	56.520
LR	74.037	70.0	71.130	70.391

For the PubMed 20K RCT dataset SVM demonstrated robust performance with an accuracy of 78.085%. However, DT showed the lowest performance among the models with an accuracy of 66.213%. The RF model outperformed the DT with better results like an accuracy of 75.298%, demonstrating the advantage of ensemble methods and NB model recorded an accuracy of 69.029%. On the other hand, LR performed comparably to SVM, with an accuracy of 77.798% showing it as a strong alternative for this dataset.

For the Heart Failure Clinical Record dataset, the SVM again showed strong results, maintaining consistent performance across the two types of medical datasets. In contrast, RF demonstrated the highest performance in this dataset with an accuracy of 78.888%, highlighting its robustness and efficiency.

These results indicate that RF, SVM and LR provide better performance on both types of datasets (Medical Record and Medical Literature). On the other hand, DT and NB exhibit performance issues, suggesting they may be less suited for these two types of datasets.

Progressing now to the concept of combining classifiers in an ensemble approach, we prove that it is highly effective in achieving a high-performing model. This strategic combination not only elevates the performance of the model but also contributes to its stability and robustness. Thus, this approach involves the individual strengths of each classifier, allowing them to complement each other and collectively make optimal decisions. The ensemble model is a strong classifier that greatly improves the overall performance and reliability of the model by taking advantage of the different features built into each individual classifier. This collaborative method not only boosts classification accuracy but also fortifies the model against potential fluctuations, providing a more resilient and dependable solution for the task at hand. To address this issue, we have employed a combination of ML methods including SVM, LR and RF classifiers. The three techniques discussed above have been combined using a soft voting classifier.

Table 4 Comparable results of different models as hybrid strategy; (a) PubMed 20K RCT (b) heart failure clinical record

	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
a				
RF-SVM	**83.123**	**83.574**	**83.326**	**83.195**
RF-LR	77.929	78.539	78.312	77.972
SVM-LR	82.278	82.453	82.325	82.302
b				
RF-SVM	80.935	80.0	80.586	80.306
RF-LR	**86.278**	**85.555**	**85.444**	**85.812**
SVM-LR	74.788	76.666	75.498	77.993

As shown in Table 4, comparing the accuracy, precision, F1-score and Recall rates of the different combinaison models that utilize a hybrid strategy with a voting classifier.

According to the results presented in Table 4, for the PubMed 20k RCT dataset, SVM-LR achieved an accuracy of 82.453%, while RF-LR achieved an accuracy of 78.539%. RF-SVM emerged with the highest accuracy of 83.574%, a notably superior performance compared to any other models, including single models targeting the same goal. This indicates that the ensemble method combining RF and SVM is particularly effective for the PubMed 20k RCT dataset.

In contrast, for the Heart Failure clinical record dataset, RF-LR demonstrated exceptional performance, achieving the highest values across all metrics: an accuracy of 85.555%, precision of 86.278%, F1-score of 85.812% and a Recall of 85.444%.

These results suggest that the combination of RF and LR is highly effective for this type of medical record data. These results highlight the importance of combined single models, while RF-SVM shines in the PubMed 20k RCT dataset and RF-LR proves to be the best choice for the Heart Failure clinical record dataset.

4.4 Ensemble Learning Models

Our primary objective was to construct a highly effective and ideal hybrid model, considering its myriad advantages. In order to do this, we examined many methodologies, including bagging and stacking, and naturally voting procedure that was previously mentioned. The analysis of bagging, stacking and voting ensemble approaches, as presented in Table 5, provides useful insights into their performance in terms of Accuracy, Precision, F1-score and Recall.

For the PubMed 20K RCT dataset, the soft voting ensemble method outperformed bagging and stacking in all evaluation metrics. Consequently, this superior

Table 5 Comparative analysis of hybridization techniques: voting, bagging and stacking; (a) PubMed 20K RCT (b) heart failure clinical record

	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
a				
Voting	83.123	83.574	83.326	83.195
Bagging	82.806	82.349	82.212	81.973
Stacking	75.860	75.874	74.914	75.866
b				
Voting	86.278	85.555	85.444	85.812
Bagging	83.1863	83.333	83.359	83.248
Stacking	83.259	82.222	82.871	82.570

performance suggests that the voting method, which combines predictions from multiple single models, effectively enhances model reliability and robustness for this dataset.

Similarly, for the Heart Failure clinical record dataset, the voting method again showed the highest performance with an accuracy of 85.555%.

This consistent superiority highlights the voting technique as a robust approach across different types of medical data.

4.5 Optimal Parameter Selection Using Genetic Algorithm (GA)

Having established that the soft voting classifier stands out as the most effective ensemble learning model, our attention turns to improving its performance through fine-tuning the weight parameter carefully and automatically. So, this next stage involves a meticulous examination and fine-tuning process, aiming to optimize the classification model settings for even greater efficiency. The weight allocated to each classifier, denoted as W, is a critical component that has a major impact on the performance of the model. This weight is determined by the mathematical equation (7) mentioned in Sect. 3.3.2. Consequently, this section will examine the impact of the weight parameter on the performance of our model. The model weights assigned in the soft voting classifier dictate the relative significance of each model in the final prediction.

By assigning uniform weights to all classifiers, each classifier will make an equal contribution to the final prediction, as depicted in Fig. 2. Nevertheless, it is important to acknowledge that certain classifiers may exhibit greater precision than others, making it less than ideal to assign them identical weights. By allocating higher weights to more accurate models, the soft voting classifier can produce more exact results and improve its overall performance.

In order to illustrate the impact of applying varying weights to the soft voting classifier, we performed tests on two types of datasets using our most optimal model.

So, it is important to note that the weight values range from 0.1 to 0.9 and their total sum is equal to 1. However, the outcomes of the previous trial are acquired by allocating uniform weights to all models in the soft voting classifier, specifically 0.5 for each (by default, i.e., we don't modify any parameters in the classifier).

In the initial phase of our experiments, it was normal to allocate a weight of 0.1 to the first classifier and 0.9 to the second classifier. This weight distribution aimed to assess the individual contributions of each classifier to the overall performance of the soft voting classifier.

For the second trial, the weights underwent a deliberate refinement process. Initially set at 0.2 and 0.8 for each classifier, we conducted subsequent adjustments, incrementing the weights to 0.3 and 0.7. Thus, this iterative process continued until we reached the final combination of weights, namely 0.9 for the first classifier and 0.1 for the second classifier.

We then evaluated the results in terms of Accuracy, Precision, F1 score and Recall to determine the optimal set of weights through this iterative process. Furthermore, assigning weights manually in each iteration of the experiment introduces several drawbacks, including the fact that manual assignment of weights lacks scalability, especially in scenarios with a large number of classifiers or complex models. As the number of classifiers increases, the task of finding an optimal set of weights becomes increasingly challenging and impractical when done manually. In contrast, employing a GA to optimize the set of weights automatically offers several advantages. A GA can explore a vast search space with high precision, which extends beyond whole numbers like 0.1, 0.2 and 0.3. Thus, a GA can assign values such as 0.84954924 and 0.15045076, offering a level of granularity that manual assignment may struggle to achieve. To sum up, while manual assignment of weight values is labor-intensive and subjective, the utilization of genetic algorithms provides a more efficient, scalable and precise method for finding the optimal set of weights.

So, this method enhances the performance of the soft voting classifier by exploring a broader search space, allowing for nuanced adjustments that may not be feasible through manual tuning.

In pursuit of optimal performance, the weights were set to 0.73644952 and 0.26355048 for SVM and RF, respectively, the GA produced impressive results in the PubMed 20k RCT dataset, this led to an accuracy of 84.133%, precision of 83.721%, F1-score of 83.712% and a Recall of 84.093%.

Furthermore, the identification of the optimal set of weights 0.84954924 for RF and 0.15045076 representing the contribution of LR, resulted in an impressive performance on the Heart Failure clinical record dataset, which indicates

that the predictions of the RF had a more substantial and influential impact on the final decision made by the ensemble learning model. This set of weights achieved an accuracy of 92.2222%, precision of 92.1436%, F1-score of 92.1687% and a Recall of 92.135%.

Table 6 illustrates the progression of classification performance from single models to a hybrid model and further to an optimized hybrid model using a GA. In the Literature Dataset, the single models (SVM, LR and RF) show moderate performance, with SVM achieving an accuracy of 78.08%, LR recording an accuracy of 77.798% and RF an accuracy of 75.298%. Hybridization significantly boosts these accuracy to approximately 83.574%, indicating the effectiveness of hybridization approach. Further weight optimization using a GA increases the accuracy to 84.133%, showcasing incremental gains from parameter fine-tuning.

In the Record Dataset, similar trends are observed with SVM, LR and RF performing variably, and hybridization elevating the accuracy to around 85.555%. The optimized hybrid model achieves a substantial performance leap, reaching an accuracy of 92.222%.

These results underscore the considerable advantages of hybridization and the further enhancements achievable through weight optimization.

Figure 4 offers a detailed breakdown of these improvements, showcasing how the models evolve in their classification when transitioning from single models to hybrid models, and further, when the weight parameters are optimized using a genetic algorithm. This comparison not only highlights the impact of hybrid strategy but also underscores the substantial performance achieved through the optimization of weights parameters (using a GA).

The improved voting classifier's ability to leverage the strengths of the individual in each separate model and generate accurate predictions exemplifies its effectiveness in ensemble learning tasks. Consequently, our proposed model yielded superior results compared to existing state-of-the-art approaches in the literature [4, 13, 32, 54], refer to Table 7.

For the PubMed 20K RCT, our proposed model achieved a precision of 83.726%, accuracy of 84.133%, F1-score of 83.712%, and a recall of 84.093%, significantly outperforming the state-of-the-art method by Anantharaman et al. [4].

Table 6 Progression of results from single models to optimized hybrid model

	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
Literature dataset				
SVM	77.826	78.085	77.932	77.845
LR	77.284	77.798	77.522	77.406
RF	74.460	75.298	75.179	74.204
Hybridization	83.123	83.574	83.326	83.195
Hybridization + optimization	83.721	84.133	84.093	83.712
Record dataset				
SVM	77.348	75.555	76.023	76.111
LR	74.037	70.000	71.130	70.391
RF	79.873	78.888	78.697	78.038
Hybridization	86.278	85.555	85.444	85.812
Hybridization + optimization	92.143	92.222	92.135	92.168

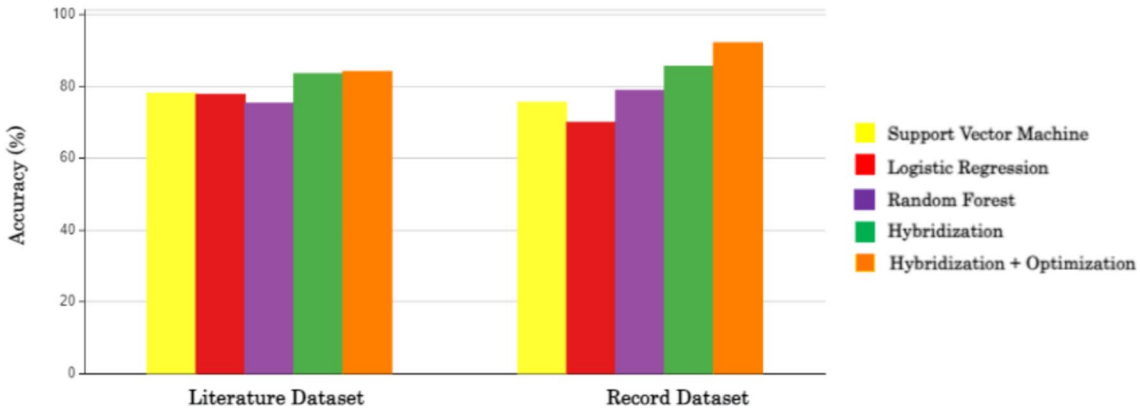

Fig. 4 Results progression

Table 7 Comparative results obtained by different literature using the same datasets

Study	Dataset	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
[4]	PubMed 20K RCT	50.5	57.5	44.6	44.4
Our approach	PubMed 20K RCT	**83.726**	**84.133**	**84.093**	**83.712**
[13]	Heart Failure Clinical Fecord	–	83.8	–	71.9
[54]	Heart Failure Clinical Record	54.0	68.79	56.25	–
[32]	Heart Failure Clinical Record	67.5	83.3	67.2	–
Our approach	Heart Failure Clinical Record	**92.143**	**92.222**	**92.135**	**92.168**

This improvement is due to the hybrid model with optimization of the weight parameter.

Similarly, for the Heart Failure clinical record dataset, our approach achieved a precision of 92.143%, accuracy of 92.222%, F1-score of 92.168% and a recall of 92.135%. This is markedly higher than the results reported by Chicco et al. [13], Firas et al. [32] and Uddin et al. [54]. The superior performance of our model on this record dataset highlights the effectiveness of our ensemble approach, particularly the optimization of weight parameters using the GA.

Overall, the experimental results validate the advantages of the hybridization approach with weight parameter optimization based on a genetic algorithm, in contrast to other works that used single models.

5 Conclusion

We have observed a rapid increase in the utilization of text classification within the healthcare sector, owing to the extensive storage of medical information in textual format. Our study aimed to develop an NLP-based medical text classification system, encompassing both types of medical text: literature and record. After cleaning the data, solve the problem of class imbalance for the record dataset with the SMOTE technique and evaluate various feature extraction techniques, including TF-IDF and CV. We conducted a comparison between single and hybrid models, followed by weight optimization using a GA, offering valuable insights into the effectiveness of our approach. Besides, a comparison among individual models such as SVM, DT, RF, NB and LR was conducted to assess their performances. Consequently, upon combining the individual classifiers, a notable improvement in classification capabilities was observed, underscoring the efficacy of hybridization. For the literature-based dataset, the best individual model achieved an accuracy of 78.085%. After hybridization and weight parameter optimization, the accuracy substantially increased to 84.133%. Similarly, for the Heart Failure Clinical Record dataset, the best single model achieved an accuracy of 78.888%. Following hybridization and weight parameter optimization, the accuracy significantly improved to 92.222%. These results underscore the superiority of our augmented hybrid model over individual traditional models. So, while each model demonstrated unique strengths and limitations, the combination of these models significantly outperformed individual models. Ultimately, our results support the idea that the ensemble learning model is a good and strong approach for text classification.

The most substantial enhancement in classification performance emerged through the optimization of weight parameters using a GA. It is valuable to consider future avenues that may contribute further to the understanding and advancements in medical text classification, utilizing diverse deep learning models to further augment efficiency.

Acknowledgements The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

Declarations

Conflict of interest Not applicable.

References

1. Acharya A (2004) GoogleScholar. https://scholar.google.com. Accessed 05 June 2024
2. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA (2017) Survival analysis of heart failure patients: a case study. PLoS One 12(7):e0181001
3. Al Amrani Y, Lazaar M, El Kadiri KE (2018) Random forest and support vector machine based hybrid approach to sentiment analysis. Proc Comput Sci 127:511–520
4. Anantharaman A, Jadiya A, Siri CTS, Adikar BN, Mohan B (2019) Performance evaluation of topic modeling algorithms for text classification. In: 2019 3rd international conference on trends in electronics and informatics (ICOEI), pp 704–708. IEEE

5. Asif M, Nishat MM, Faisal F, Dip RR, Udoy MH, Shikder M, Ahsan R et al (2021) Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease. Eng Lett 29(2):731–741
6. Basarkar A (2017) Document classification using machine learning
7. Ben Abdennour G, Gasmi K, Ejbali R (2023) Ensemble learning model for medical text classification. In: International conference on web information systems engineering, pp 3–12
8. Bhavani A, Kumar BS (2021) A review of state art of text classification algorithms. In: 2021 5th International conference on computing methodologies and communication (ICCMC), pp 1484–1490. IEEE
9. Blagus R, Lusa L (2015) Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinform 16(1):1–10
10. Chadaga K, Chakraborty C, Prabhu S, Umakanth S, Bhat V, Sampathila N (2022) Clinical and laboratory approach to diagnose COVID-19 using machine learning. Interdiscip Sci: Comput Life Sci 14(2):452–470
11. Chang V, Bailey J, Xu QA, Sun Z (2023) Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput Appl 35(22):16157–16173
12. Charbuty B, Abdulazeez A (2021) Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends 2(01):20–28
13. Chicco D, Jurman G (2020) Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20(1):1–16
14. Dernoncourt F, Lee JY (2017) Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071
15. Dernoncourt and Lee.:PubMed. https://pubmed.ncbi.nlm.nih.gov/. Accessed 05 June 2024
16. Dong X, Yu Z, Cao W, Shi Y, Ma Q (2020) A survey on ensemble learning. Front Comput Sci 14:241–258
17. Du J, Rong J, Wang H, Zhang Y (2021) Neighbor-aware review helpfulness prediction. Decis Support Syst 148:113581
18. Fernández A, Garcia S, Herrera F, Chawla NV (2018) SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 61:863–905
19. Ge Y-F, Bertino E, Wang H, Cao J, Zhang Y (2023) Distributed cooperative coevolution of data publishing privacy and transparency. ACM Trans Knowl Discov Data 18(1):1–23
20. Ge Y-F, Orlowska M, Cao J, Wang H, Zhang Y (2022) MDDE: multitasking distributed differential evolution for privacy-preserving database fragmentation. VLDB J 31(5):957–975
21. Ge Y-F, Wang H, Cao J, Zhang Y (2022) An information-driven genetic algorithm for privacy-preserving data publishing. In: International conference on web information systems engineering, pp 340-354
22. Ge Y-F, Yu W-J, Cao J, Wang H, Zhan Z-H, Zhang Y, Zhang J (2020) Distributed memetic algorithm for outsourced database fragmentation. IEEE Trans Cybern 51(10):4808–4821
23. Grandini M, Bagli E, Visani G (2020) Metrics for Multi-Class Classification: an Overview. arXiv preprint arXiv:2008.05756
24. Heart failure clinical records. UCI Machine Learning Repository (2020). https://doi.org/10.24432/C5Z89R
25. Heredia B, Khoshgoftaar TM, Prusa J, Crawford M (2016) An investigation of ensemble techniques for detection of spam reviews. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA), pp 127–133. IEEE
26. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W (2018) Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genom Proteom 15(1):41–51
27. Hussain L, Lone KJ, Awan IA, Abbasi AA, J-u-R P (2022) Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. Waves Random Complex Media 3:1079–1102
28. Immanuel Savio D, Chakraborty UK (2019) Genetic algorithm: an approach on optimization. In: 2019 international conference on communication and electronics systems (ICCES), pp 701–708. https://doi.org/10.1109/ICCES45898.2019.9002372
29. Jinha AE (2010) Article 50 million: an estimate of the number of scholarly articles in existence. Learn Publ 23(3):258–263
30. Kadhim AI (2019) Survey on supervised machine learning techniques for automatic text classification. Artif Intell Rev 52(1):273–292
31. Kambar MEZN, Nahed P, Cacho JRF, Lee G, Cummings J, Taghva K (2022) Clinical text classification of Alzheimer's drugs' mechanism of action. In: Proceedings of sixth international congress on information and communication technology: ICICT 2021, London, Vol 1, pp 513–521. Springer
32. Ketata F, Al Masry Z, Zerhouni N, Yacoub S (2023) Explainable machine learning approach with augmentation for mortality prediction. In: 2023 IEEE international conference on advanced systems and emergent technologies (IC_ASET). IEEE
33. Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J (2022) An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors 22(19):7268
34. Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cognit Comput Eng 2:40–46
35. Laakso M, Welling P, Bukvova H, Nyman L, Björk BC, Hedlund T (2011) The development of open access journal publishing from 1993 to 2009. PloS One 6(6):e20961
36. Lambora Annu; Gupta, Kunal; Chopra, Kriti (2019) Genetic algorithm—a literature review. In: 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), pp 380–384. https://doi.org/10.1109/COMITCon.2019.8862255
37. Larsen P, Von Ins M (2010) The rate of growth in scientific publication and the decline in coverage provided by science citation index. Scientometrics 84(3):575–603
38. Luo X (2021) Efficient English text classification using selected machine learning techniques. Alex Eng J 60(3):3401–3409
39. MEDLINE (2024) https://www.nlm.nih.gov/databases/databases_medline.html. Accessed 05 June
40. Mercadier Y (2020) Classification automatique de textes par réseaux de neurones profonds: application au domaine de la santé. Université Montpellier
41. Naeem MZ, Rustam F, Mehmood A, Ashraf I, Choi GS et al (2022) Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. PeerJ Comput Sci 8:e914
42. Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya INH, Ismail M (2021) SMOTE for handling imbalanced data problem: a review. In: 2021 sixth international conference on informatics and computing (ICIC), pp 1–8. IEEE
43. Prabhat A, Khullar V (2017) Sentiment classification on big data using Naïve Bayes and logistic regression. In: 2017 international conference on computer communication and informatics (ICCCI), pp 1–5. IEEE
44. Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P (2023) Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Syst Appl 212:118715
45. Raychaudhuri K, Kumar M, Bhanu S (2017) A comparative study and performance analysis of classification techniques: support

vector machine, neural networks and decision trees. In: Advances in computing and data sciences (ICACDS), pp 13–21. Springer
46. Rish I et al (2001) An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. vol 3(22), pp 41–46
47. Rustam F, Saher N, Mehmood A, Lee E, Washington S, Ashraf I (2023) Detecting ham and spam emails using feature union and supervised machine learning models. Multimed Tools Appl 82(17):1–17
48. Sakkis G, Androutsopoulos I, Paliouras G, Karkaletsis V, Spyropoulos CD, Stamatopoulos P (2001) Stacking classifiers for anti-spam filtering of e-mail. arXiv preprint cs/0106040
49. Shah K, Patel H, Sanghvi D, Shah M (2020) A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augment Hum Res 5:1–16
50. Sohail A (2023) Genetic algorithms in the fields of artificial intelligence and data sciences. Ann Data Sci 10(4):1007–1018
51. Sutton CD (2005) Classification and regression trees, bagging, and boosting. Handb Stat 24:303–329
52. Tripathy A, Anand A, Rath SK (2017) Document-level sentiment classification using hybrid machine learning approach. Knowl Inf Syst 53:805–831
53. Uddin MJ, Ahamad MM, Sarker PK, Aktar S, Alotaibi N, Alyami SA, Kabir MA, Moni MA (2023) An integrated statistical and clinically applicable machine learning framework for the detection of autism spectrum disorder. Computers 12(5):92
54. Uddin S, Haque I, Lu H, Moni MA, Gide E (2022) Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep 12:6256
55. Utomo MRA, Sibaroni Y (2019) Text classification of British English and American English using support vector machine. In: 2019 7th international conference on information and communication technology (ICoICT), pp 1–6. IEEE
56. Yang F-J (2018) An implementation of naive Bayes classifier. In: International conference on computational science and computational intelligence (CSCI), pp 301–306. IEEE
57. Yin J, Tang M, Cao J, You M, Wang H, Alazab M (2022) Knowledge-driven cybersecurity intelligence: software vulnerability coexploitation behavior discovery. IEEE Trans Ind Inform 19(4):5593–5601
58. You M, Yin J, Wang H, Cao J, Wang K, Miao Y, Bertino E (2023) A knowledge graph empowered online learning framework for access control decision-making. World Wide Web 26(2):827–848