Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240518

Comparison of Ensemble Learning Methods for Classification in Cancer Registries

Nico SCHULT[a,1], Timo WOLTERS[a], Marc HERMES[a], Klaas DÄHLMANN[a] and Andreas HEIN[b]
[a] *Division Health, OFFIS - Institute for Information Technology, Escherweg 2, Oldenburg, Germany*
[b] *Department of Health Services Research, Carl von Ossietzky University Oldenburg, Oldenburg, Germany*

Abstract. Significant developments are currently underway in the field of cancer research, particularly in Germany, regarding cancer registration and the use of medical information systems. The use of such systems contributes significantly to quality assurance and increased efficiency in data evaluation. The growing importance of artificial intelligence (AI) in cancer research is evident as these systems integrate AI for various purposes, i.e. to assist users in data analysis. This paper uses ensemble learning to classify the graphical user interface state of the medical information system CARESS. The results show that all ensemble learning models utilized achieved good performance. In particular, the gradient boosting algorithm performed the best with an accuracy of 97%. The results represent a starting point for further development of ensemble learning in medical data analysis, with the potential for integration into various applications such as recommender systems.

Keywords. artificial intelligence in medicine, ensemble learning, hyperparameter optimization, gradient boosting, adaboost, random forest, cancer registry

1. Introduction

Research in the field of cancer care is currently undergoing development in Germany, particularly in the context of cancer registration and the associated use of medical information systems. A cancer registry is an organization that systematically collects, stores, analyses, interprets and publishes information on tumors and their treatment from hospitals, general practitioners and oncology centers [1].

The central role of modern medical information systems in cancer registries is essential to ensure data quality and increase the efficiency of data analyses. These systems contribute significantly to the quality assurance and validity of research data and form a pillar for advances in cancer care [2].

Medical data inherently presents a number of challenges due to its complexity, which has an impact on the preparation of analyses. As this complexity often overwhelms

[1] Corresponding Author: Nico Schult, OFFIS - Institute for Information Technology, Escherweg 2, Oldenburg, Germany; E-mail: nico.schult@offis.de.

users, a solution that provides potentially missing information in particular is required [2].

One possible solution could be the implementation of a recommender system. Such a system addresses the complexity by providing personalized help. This simplifies the onboarding process for new users and facilitates efficient navigation through the system. The first step of providing a recommender system to the user is to recognize and analyze the user's activities and interactions in the software [3].

Due to the comprehensive nature of data analysis, medical information systems often have complex and extensive graphical user interfaces (GUI) that include various functions and modules. A GUI state refers to the specific configuration and appearance of elements within a graphical interface at a given moment. The GUI encapsulates the current visual and functional representation of an application or software, reflecting user interactions and system responses. Determining the exact state of the GUI based purely on programmatic logic can be difficult.

One approach for recognizing the GUI state is the application of artificial intelligence (AI). Medical information systems already integrate AI for various purposes, including the assistance of users in analyzing data [4]. Machine learning models require a lot of data, but in the medical domain, collecting a sufficient amount of data is often a challenge. In addition, machine learning models tend to overfit when applied to such limited the amounts of data.

For this reason, we will take a closer look at ensemble learning. Ensemble learning is a machine learning method in which multiple models are combined with the goal to improve classification accuracy and robustness. This method makes it possible to integrate different learned models to create a compact model [5]. For medical data in particular, Srivastava et al. [6] have shown that ensemble learning methods are better suited than conventional machine learning methods.

Therefore, in the context of analyzing cancer registry data, ensemble learning offers a promising solution to overcome the challenges posed by the complexity and sensitivity of the data and the structure of cancer registries. Ensemble learning methods allow not only the replacement of learned models, but also the combination of these models to perform a more accurate analysis. As such, different methods of ensemble learning exist, and it needs to be determined which one is best suited to classify the current GUI state in a medical information system for cancer registers.

In this paper, we aim to classify the current GUI state of the medical information system CARESS using Ensemble Learning. Contributions of this work include:

- Comparison of ensemble learning methods suitable for classification, along with the use of different metrics to evaluate the performance of the models.
- Examination of performance improvement of ensemble learning models through hyperparameter optimization.

The remainder of this paper is structured as follows: In the next section, the methodology is examined in detail. The focus is on the individual steps to generate an ensemble learning model to classify the current GUI state. The results of the trained ensemble learning models are then presented in section 3. To evaluate the best model, the results are interpreted and discussed in section 4. An outlook for further work is also provided in this section. Finally, the section 5 summarizes the paper.

2. Methods

As mentioned above, Srivastava et al. [6] compares the performance of ensemble learning versus conventional machine learning models and successfully demonstrated that ensemble learning is superior to machine learning models in classifying medical data. Among the various ensemble learning methods investigated, Random Forest and AdaBoost proved to be good methods for their use case. For this reason, we also use these two methods for our use case. Additionally, our own preparatory work for this paper has shown that gradient boosting appears to be also a promising method. In the following section, we will take a closer look at the implementation and training of these methods to create an ensemble learning model to classify a GUI state of a medical information system.

Random Forest works by training numerous decision trees and making the prediction by majority voting on all trees. This increases the robustness of the model and reduces the risk of overfitting when compared to a single tree [5]. AdaBoost emphasizes the weighting of errors by training weak classifiers one by one, placing more weight on misclassified instances. This process enables the creation of an improved model, which is built with emphasis on the errors of the prior models [7]. Gradient Boosting builds an ensemble of decision trees sequentially, with each tree aiming to correct the errors of the previous ones. It optimizes a loss function by fitting each new tree to the residuals of the combined ensemble, resulting in a model with enhanced predictive accuracy [7].

The ensemble learning models are trained using supervised learning, which requires labeled data. The data was obtained from a study where different users had to create analyses in CARESS and can be found at Harvard Dataverse under the title Application of the clinical model with existing systems [8]. CARESS is a medical information system that is used in many cancer registries in Germany[2]. The dataset consists of a large number of extracted display events. A display event contains all information about the user interface, such as the position and text of a button element. This information must be specifically pre-processed in order to make it optimally usable for training. The available dataset contains a set of 12 different labels, with each label representing a unique GUI state. The data is pre-processed in such a way that only categorical features are extracted.

Following this, the Countvectorizer method is used. The Countvectorizer method captures the frequency of extracted text features in a particular display event texts compared to the entire data collection [9]. This vectorization creates a numerical representation of the text data, which makes it possible for the ensemble learning model to capture semantic similarities between different display events. This step is crucial in order to create the basis for precise and meaningful models in the context of data analysis. After defining the ensemble learning methods and describing the structure of the data, the following step explains the modeling process in more detail.

First, the data set was divided into training and test data in a 70/30 ratio. The models were then trained using the training data and evaluated using the test data set. Then metrics were introduced for the evaluations, such as Precision, Recall, F1-Score and Accuracy. These metrics are used to comprehensively evaluate the performance of the ensemble models. Precision measures the accuracy of positive predictions, Recall measures the sensitivity to true positives, and F1-Score measures the harmonic mean combination of Precision and Recall. Accuracy is a measure of how well a model handles the overall number of correct predictions in relation to the number of data points [10].

[2] Lower Saxony, Hamburg, Bremen, Baden-Württemberg, Thuringia.

After evaluating the models, the next step is to optimize them. A proven approach for optimizing the models is the use of GridSearchCV (Grid Search Cross-Validation). This method systematically searches through different hyperparameter combinations to find the best possible settings for a model [11]. The best parameter combinations are then added to the model and re-evaluated to measure their performance.

3. Results

After explaining the steps to create an ensemble learning model for the given problem in section 2, this section is dedicated to analyzing the results to evaluate an optimal model. Table 1 presents the performance metrics of three machine learning models, namely Random Forest, AdaBoost, and Gradient Boosting, as well as their optimized versions marked with (*). The metrics evaluated are accuracy, precision, recall, and F1 score, each calculated as a weighted average considering the number of observed labels.

The results show the effectiveness of the optimization process, as the optimized Random Forest model shows improvements in all metrics compared to its non-optimized counterpart. Similarly, the AdaBoost algorithm shows improved performance after optimization. The Gradient Boosting algorithm shows the highest improvement with optimization and achieves the best overall results for each metric measured.

Table 1. Performance Metrics for Different Models (optimized models marked with (*))

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	94.3	93.9	93.1	93.5
AdaBoost	92.6	91	91.5	91.1
Gradient Boosting	94.3	93.3	93.3	93.1
Random Forest (*)	94.4	94.4	94.3	94.4
AdaBoost (*)	93.6	92.6	92.7	92.6
Gradient Boosting (*)	97	94.8	97.1	95.8

4. Discussion

The non-optimized Random Forest model already performs well, but the optimization process leads to slight enhancements in all metrics, making it a reliable choice. AdaBoost, even in its non-optimized form, shows comparable results, and optimization further refines its performance. The two optimized models achieved values between 92.6 and 94.4 in all metrics.

However, it is the Gradient Boosting algorithm that stands out, both in its non-optimized and optimized versions. This model consistently achieves the highest scores across all metrics, highlighting its suitability for classifying GUI states in the context of medical information systems.

The optimized versions of the models show notable improvements across all evaluated metrics compared to their non-optimized counterparts. This emphasizes the importance of fine-tuning hyperparameters to maximize the potential of ensemble learning models. An accuracy of 97% indicates that almost all instances were correctly

classified by the Gradient Boosting model, highlighting its ability to accurately classify the GUI state. Overall, it can be concluded that all ensemble learning models utilized achieved excellent results. The optimized models appear to be suitable for reliably classifying the GUI state within medical information systems.

5. Conclusions

In summary, we have successfully shown that ensemble learning, in particular through the application of Gradient Boosting, is a promising approach for classifying the GUI state in medical information systems. The results of the comprehensive evaluation of the models based on commonly used metrics such as accuracy, precision, recall and F1-score underlines the effectiveness of ensemble learning in a medical context.

The presented approach offers a basis to implement the first step of a recommender system as described in [3]. The results emphasize the effectiveness of hyperparameter optimization in enhancing the performance of ensemble learning models for GUI state classification.

Future work may build upon this research and implement the next steps to develop a recommender system for a medical information system.

References

[1] Bundesanzeiger. Krebsfrüherkennungs- und -registergesetz (KFRG) [Internet]. Bundesgesetzblatt; 2013 Apr 3 [cited 2024 May 13]. Available from: https://www.bgbl.de/xaver/bgbl/start.xav?start=//*%5B@attr_id=%27bgbl113s0617.pdf%27%5D#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl113s0617.pdf%27%5D__1714122234116.

[2] Bianconi F, Brunori V, Valigi P, La Rosa F, Stracci F. Information Technology as Tools for Cancer Registry and Regional Cancer Network Integration. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 2012 Oct;42(6):1410-1424. doi: 10.1109/TSMCA.2012.2210209.

[3] Michael J, Rumpe B, Varga S. Human Behavior, Goals and Model-Driven Software Engineering for Assistive Systems. In: Koschmider A, Michael J, Thalheim B, editors. EMISIA Workshop 2020; 2020 May 14-15; Kiel, Germany. p. 11-18.

[4] Combi C, Pozzi G. Clinical Information Systems and Artificial Intelligence: Recent Research Trends. Yearbook of medical informatics. 2019 Sep;28(01):083-094. doi: 10.1055/s-0039-1677915.

[5] Mienye ID, Sun Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access. 2022 Sep;10:99129-99149. doi: 10.1109/ACCESS.2022.3207287.

[6] Srivastava S, Yadav RK, Narayan V, Mall PK. An Ensemble Learning Approach For Chronic Kidney Disease Classification. Journal of Pharmaceutical Negative Results. 2022 Dec;13(10):2401-2409. doi: 10.47750/pnr.2022.13.S10.279.

[7] Sagi O, Rokach L. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery. 2018 Feb;8(4):e1249. doi:10.1002/widm.1249.

[8] Wolters T. Application of the clinical model with existing systems [Internet]. Harvard Dataverse; 2024 Mar 14 [cited 2024 May 13]. doi: 10.7910/DVN/RPGLHX. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/RPGLHX.

[9] Turki T, Sekhar S. Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer. Applied Sciences. 2022 Jun;12(13):6611. doi: 10.3390/app12136611.

[10] Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process. 2015 Mar;5(2):1-11. doi: 10.5121/ijdkp.2015.5201.

[11] Ranjan GSK, Verma AK, Radhika S. K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. In: 5th International Conference for Convergence in Technology 2019; 2019 Mar 29-31; Pune, India. doi: 10.1109/I2CT45611.2019.