Comparative Analysis of Machine Learning Techniques and Physical-Chemical Approaches for Celiac Disease Diagnosis: Evaluating Predictive Performance and Feature Importance

Ameya Jinnuri & Maheshwar Murugesan

Received August 01, 2024
Accepted October 15, 2024
Electronic access November 15, 2024

This study presents a comparative analysis of ML models and traditional physical-chemical diagnostic methods for celiac disease, emphasizing the predictive advantages of ML in reducing diagnostic invasiveness and enhancing early intervention opportunities. Variables in the expression of symptoms and overlapping symptoms with other gastrointestinal disorders make early diagnosis very hard. For this study, advantages will be accrued from open data repositories using multiple datasets to develop a more robust machine-learning model that can accurately predict. It uses random forest, gradient boosting, support vector classifier, and logistic regression algorithms. The important features in this work were age, diabetes, and abdominal pain. The paper is very good in addressing overfitting, with key considerations regarding feature importance and careful cleaning of data that generally will help generalizability and applicability in the clinic. Results show how this may reduce the need for invasive procedures and facilitate very early dietary interventions to improve outcomes of patients, optimizing healthcare resources.

Analyzing Determinants of Medical Costs and Predictive Modeling Using Machine Learning Techniques

Celiac disease is characterized by an immune response triggered by the ingestion of gliadin, a component of gluten. This leads to the production of antibodies against tissue transglutaminase (tTG), causing damage to the small intestine, specifically villous atrophy, which is central to the pathology of celiac disease[1]. Celiac disease presents with a wide range of gastrointestinal symptoms, such as diarrhea, bloating, and non-gastrointestinal symptoms like fatigue, dermatitis herpetiformis, and neurological issues. These symptoms overlap with disorders like irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD), complicating the diagnostic process. Timely and proper diagnosis makes a difference for long-term health complications like malnutrition, osteoporosis, and increased risk for certain malignancies[2] Current diagnostic methods for celiac disease, such as serological testing (tTG-IgA), endoscopy, and biopsy, often require invasive procedures and may fail to detect the disease in early stages. While physical-chemical approaches rely heavily on invasive procedures such as endoscopy and biopsy, ML technologies provide a non-invasive alternative by analyzing large datasets, including serological markers and genetic data. The key distinction lies in machine learning's ability to process complex variables and detect subtle patterns, potentially bypassing the need for traditional methods that often miss early-stage diagnoses.

Machine learning has accelerated change in healthcare from the analysis of imaging through predictive analytics. Machine learning models will find complex relationships and patterns in biomarkers, imaging data, and clinical presentations that might elude traditional diagnostic methods[3,4] ([5,6]. Celiac disease is one of the cases in which ML models are highly effective at evaluating demographic data, genes, and clinical symptoms of the patient to provide a better prediction of the onset of the disease, hence facilitating effective and timely medical intervention[7].

Open data repositories have grown dramatically, greatly facilitating the process of developing ML applications in health care. These large, high-quality data sets are central to the training and validation of ML models. In the case of celiac disease, many of these datasets drawn from open repositories, particularly the National Institutes of Health, European Bioinformatics Institute, and other healthcare databases, are tremendously rich in sources of demographic data of patients, genetic markers, and clinical case histories. Using dataset sources like these to train robust machine learning models that generalize across diverse populations is possible.

The impact of this work is that it really will change how the diagnosis of celiac disease is made. An ML model predicting the potential presence of celiac disease in a patient through demographics and other relevant factors may have some critical advantages. First, improving diagnostic accuracy is achievable by reducing dependence on invasive procedures like endoscopy and biopsy, enabling the possibility of an early diagnosis with

minimal invasion. This will have the impact of improving comfort for the patient while at the same time-saving healthcare costs that arise as a result of unnecessary procedures. Early identification of those at risk allows timely dietary interventions and the formulation of a care plan aimed at reducing the incidence of serious and classic complications of celiac disease. The introduction of a gluten-free diet very early may prevent long-term damage that undiagnosed celiac disease may cause to the patient. Finally, healthcare systems can use this information to optimize resource allocation to people who are high-risk and require closer monitoring and intervention. This enables more efficient use of health resources and tends to bring about better patient outcomes. This study can also, not in the last turn, add valuable data for the epidemiology of celiac disease and therefore to public health strategies related to prevention and early detection. Demographic data such as age, gender, and geographical location were combined with genetic markers like HLA-DQ2/DQ8 to enhance the predictive power of the model.

The overall objective of the study is to develop a machine-learning model for the accurate prediction of celiac disease likelihood using open repository datasets. It picks up trends and risk factors for diseases by amalgamating patient demographics, genetic information, and clinical history. Its secondary objectives are the validation of the model in different datasets, assessment of its generalizability, and investigation of integration into clinical practice. It looks to the future when machine learning combined with open data will empower early diagnosis and improve celiac disease management. If this model is implemented appropriately, it would greatly improve patient outcomes and hence be a huge contribution to predictive healthcare analytics.

The gap that is being addressed in this study is the lack of non-invasive procedures in the diagnosis of celiac disease. A majority of procedures for the diagnosis of celiac disease are invasive, such as the endoscopy or biopsy of the small intestine. This study acts as one of the few stepping stones for non-invasive procedures. With the rise of computing power and development of machine learning and AI, this study is one of the pioneering studies in the field of predicting diagnoses for celiac disease.

Literature Review

In this literature review, the evolving landscape of predictive modeling in diagnosing celiac disease using machine learning (ML) and artificial intelligence (AI) is explored. This review aims to analyze the implications of utilizing ML and AI solutions in the medical field, focusing on enhancing diagnostic accuracy and patient care through advanced data analytics.

Predictive Modeling of Celiac Disease Using Machine Learning

Machine learning is being inculcated into lines multiple to human health, including, but not limited to, the diagnosis of celiac disease. Celiac disease is an autoimmune disorder mediated by intestinal ingestion of gluten and causes damage in the small intestine. Its early and accurate diagnosis is very important to prevent long-term complications of the disease, such as malnutrition, osteoporosis, and certain malignancies. Traditional physical-chemical diagnostic approaches, while reliable for advanced stages of celiac disease, fall short in early detection due to their reliance on visible intestinal damage or elevated antibodies. In contrast, ML models excel at identifying complex, non-linear patterns in data, including early-stage markers that physical-chemical methods may overlook. This highlights a significant shift toward data-driven, predictive healthcare. Machine learning in most aspects has been intrinsic to healthcare applications, from image analysis to predictive analytics. Zhou et al. demonstrated that random forests and support vector machines can achieve an accuracy of 93% in predicting celiac disease using EHR-based data, significantly outperforming logistic regression at 85%. Their study highlights the potential of tree-based models in clinical settings for early diagnosis[8]. In this respect, their study showed tree-based models, particularly random forests and support vector machines, to be highly accurate in predicting the disease, underpinning the effectiveness of those models within a clinical setting.[8]. Smith and Johnson reviewed the application of machine learning in the diagnosis of celiac disease in 2019. The strongest models were random forests and gradient boosting according to them[9]. This putting all research into a very exhaustive framework of what is known so far may be of importance to understanding the strengths and limitations of different ML techniques.[9].

Enhancing Diagnostic Accuracy with Machine Learning

The integration of genetic data with demographic information in existing literature is seen to generally improve diagnostic accuracy. For example, Garcia and Martinez used genetic markers HLA-DQ2 and HLA-DQ8 in combination with clinical data to enhance the specificity and sensitivity of their predictive models[6]. By applying decision trees and random forests, they achieved a prediction sensitivity of 92%, demonstrating the importance of integrating genetic profiles in diagnostic models.. Accordingly, their findings support our holistic data integration approach by underlining the importance of different data types within predictive modeling.[6]. Kumar et al. pointed out the feature's importance in terms of predictive analytics for celiac disease[10]. As was shown, clinical history and genetic markers are the two major driving variables for model predictions, hence further justifying the appropriateness of chosen features.

Their paper discussed one of the issues with overfitting and the relevance of robust data preprocessing, which is very important in developing reliable ML models.[10].

Advances in Early Detection and Predictive Analytics

Early detection of celiac disease may significantly improve patient outcomes and healthcare costs. ee and Kim provided 2022 explorations on the use of deep learning models in detecting the early stages of the disease and proved that compared to traditional machine learning techniques, these models yield improved performance[11]. Our study focuses on machine learning, and these findings could only reveal that improved algorithms open up new opportunities to improve early diagnostic capabilities.[11]. Patel and Singh (2019) studied the role that ML plays in predicting autoimmune diseases, including celiac disease. Their work added a broader perspective on the application domains of ML in the diagnostics of autoimmune diseases and further emphasized the need for integrating demographic data with clinical data in any predictive model.[12].

Practical Implications and Future Directions

This comparative study underscores the practical superiority of ML models over traditional diagnostic methods. Physical-chemical approaches, while useful, often necessitate invasive testing that may only detect celiac disease in its later stages. ML offers a paradigm shift, enabling non-invasive early detection by identifying patterns in demographic and serological data. This not only improves diagnostic accuracy but also reduces the need for costly and uncomfortable procedures. Robinson and Taylor showed the overall high predictive accuracy for ensemble methods like random forest and gradient boosting, which came incidentally to our models used in this study[13]. Their study in 2021 supports these techniques in clinical practice to help improve diagnostic precision.[13]. In a 2022 paper, Davis and Brown reviewed ML approaches for the prediction of autoimmune diseases like celiac disease[14]. Their results underline once again the necessity of baseline models like logistic regression and support vector classifiers, which need to form part of this study for benchmarking in the validation of predictive performance.[14] Williams and Jones focused on the integration of clinical data with ML techniques, thus placing a premium on data quality and preprocessing[15]. Their research, therefore, attaches prime importance to treating the data with utmost care to increase model accuracy and reliability—two very major aspects of our study.[15] Chen and Wang used demographic data such as age, gender, and family history of autoimmune disorders; laboratory results including tissue transglutaminase antibodies (tTG-IgA), total serum IgA levels, and iron deficiency markers; and clinical symptoms such as chronic diarrhea, abdominal pain, and unexplained weight loss from electronic health records (EHR) to predict celiac disease[16]. Random forest and gradient boosting outperformed logistic regression and support vector machines due to their ability to handle non-linear relationships and interactions between multiple variables in the dataset. They found that random forests and gradient boosting performed best, validating the choice of these models in our research and reinforcing the potential of ML in improving early diagnosis and patient care[16]. Machine learning has improved the speed and accuracy of celiac disease diagnosis by analyzing large datasets with complex variables like genetic and clinical data, which would be difficult to process using traditional diagnostic techniques[16]. Diagnosis should be accurate and early to prevent long-term complications of health such as malnutrition, osteoporosis, and certain malignancies[2]. Machine learning models provide a very powerful tool for the analysis of complex data sets, identification of patterns, and diagnostic accuracy beyond conventional techniques.

It has been applied to all aspects of healthcare ranging from image analysis to predictive analytics. Zhou et al. have referred to the application of logistic regression, random forests, and support vector machines in predicting celiac disease by data from electronic health records[8]. The authors of the research demonstrated brilliant accuracies of tree-based models in predicting the disease, mainly random-forest and support-vector machines, underlining their effectiveness in the clinical setting. (Zhou et al., 2020). Smith and Johnson performed a systematic review of the application domains of machine learning in the diagnosis of celiac disease in 2019[9]. Smith and Johnson mentioned that random forests and gradient boosting took first and second place, respectively, as the most powerful models in that domain, therefore giving a completely realistic picture of the current research landscape. This review would have mainly a single influence on understanding the potentials and limitations of these multiple ML techniques, all of which help in placing any new studies within already existing research frameworks[9].

Demographic data has been combined with these models and diagnostic accuracy enhanced. Much of the work seems to be focused on how one combines demographic data with genetic data in the combination process.Garcia and Martinez commented regarding how genetic profiling, which is combined with clinical data, might lead to specificity and sensitivity in the predictions for celiac disease[6]. This justifies our holistic data integration approach, combining demographic, genetic, and clinical data to enhance predictive power.[6]. Kumar et al. paid much attention to feature importance in predictive analytics for celiac disease[10]. They represented that the largest share in predictions by a model is contributed to by clinical history and genetic markers, which raises confirmations of the features selected for our study. Their paper has further highlighted issues related to overfitting and the importance of robust data preprocessing, which are critical in developing reliable ML models.[10]

Obviously, this could eventually affect patient outcomes and healthcare expenditure for people with celiac disease. Lee and Kim examined the early diagnosis of this disorder by training deep learning models to show that this method is far better than the conventional techniques used in ML[11]. As this paper is oriented toward machine learning, their results have to show that newer algorithms definitely can do a better job and therefore suggest further ways to improve the early diagnostic capabilities[11]. Patel and Singh, in their paper, have tried to look into what role machine learning can play in predicting autoimmune diseases like celiac disease[12]. It could contribute to understanding the role played by applying ML in autoimmune disease diagnostics about underlying demographic and clinical data integration in the creation of predictive models.[12]. This makes the practical application of machine learning in diagnosis very important for celiac disease. In a 2021 study, Robinson and Taylor found that some ensemble methods similar to random forests and gradient boosting used in this research had a high predictive accuracy. Robinson and Taylor found that random forests and gradient boosting increased diagnostic accuracy by 15% compared to conventional diagnostic methods in clinical practice, demonstrating their utility for celiac disease detection[13]. Comparative work in the prediction of autoimmune diseases using ML approaches was carried out by Davis and Brown in their 2022 paper. Their results underscore the baseline value of models like logistic regression and support vector classifiers used here for benchmark setting and validation of predictive performance[14].

Williams and Jones underlined how ML techniques are considered to address clinical data integration, the role of data quality, and preprocessing[15]. Their finding underscores that scrupulous treatment of data would enhance model accuracy and reliability—the two most important parameters in our present study[15]. In 2021, Chen and Wang applied EHR with machine learning models in the risk prediction for celiac disease. In their research, Random Forest and Gradient Boosting were the best-performing algorithms, thus further ascertaining this model choice of our study and showing the potential of ML in diagnosis improvement and patient care[16].

Although ML has shown promise in celiac disease diagnosis, limitations remain, such as the lack of external validation in many studies and challenges in applying genetic data across different populations. It identifies previous studies having used different machine learning models to attain better diagnosis accuracy and notes that tree-based machine learning models, especially random forests and gradient boosting, yield good results for the same. This review also illustrates that in the future, heterogeneous data sources—from genotype profiles to clinical histories—will be included to further drive model specificity and sensitivity. Nevertheless, overfitting and the requirement for robust data preprocessing remain very critical challenges. This insight is leveraged to construct a predictive model for celiac disease from open repository datasets, with an added focus on demographic, genetic, and clinical data integration. The focus of this paper is on enhancing generalizability and applicability in machine learning models in the clinical setting by answering identified challenges, such as overfitting and feature selection. To some degree, this may be considered rather novel in light of the holistic integration of different data types and the use of advanced preprocessing techniques that might help alleviate common problems of predictive modeling. It is for this reason that the work, therefore, will enhance the quality of treatment given to patients at risk of acquiring celiac disease and hence affect an improved early diagnosis, thereby affecting the whole field of predictive healthcare analytics.

Materials & Methods

Materials

The materials utilized in this study are an 8-core M1 processor with 16 GB of RAM for data preprocessing, cleaning, and hyperparameter tuning. Additionally, Anaconda open ecosystem and Jupyter notebooks were utilized to record and write code in Python. Additionally, various Python libraries from the Anaconda open ecosystem were used, including NumPy, Sci-kit Learn, Matplotlib, and Seaborn. Sci-kit Learn is a library that contains all of the machine learning models that were used within this study, and NumPy was used for mathematical calculations, like the accuracy, precision, and F1 scores of the various models. Matplotlib and Seaborn are statistical modeling libraries that were used for statistical visualization utilizing things such as confusion matrices.

Methods

In this section, the methods used to determine the most important factors in celiac disease diagnosis and the methods by which the best model for predicting celiac disease is determined. This section will also discuss ways that the models were enhanced and issues were avoided. Additionally, statistical measures such as Chi-Square tests and p-values will be determined to determine the significance of the classifications that are made by the machine learning models.

Feature Importance

Feature importance in ML models, such as Random Forest and Gradient Boosting, stands in contrast to the reliance on visible markers in traditional physical-chemical diagnostics. ML models rank features like IgA, IgM, and abdominal pain in a predictive hierarchy, allowing for nuanced diagnoses that physical-chemical methods, dependent on macroscopic tissue damage, may miss. Feature importance ranks the most important

features/inputs and their importance on the output. This study used a dataset of 1,200 patients, including 14 features such as age, gender, and clinical history. After preprocessing, the dataset contained 1,000 instances with 9 features after the removal of redundant or overfitting columns. The dimensions of the dataset were 1,000 rows and 9 columns: Age, Gender, Diabetes, Diabetes Type, Diarrhea, Abdominal pain, Short stature, Sticky stool, Weight loss, IgA, IgG, IgM, and Marsh (intestinal damage), and celiac disease type. However, 5 out of the 14 columns were dropped, as they were causing major issues within the models and the functionality of the models.

Overfitting

Overfitting was an issue that was encountered at the beginning of this study, where the models would have learned far too well on the data that they were trained on. This is an issue because the models have extremely poor performance on new instances, and using new instances to generate predictions is the purpose of the study. This comes from some of the columns where it is far too easy to generate predictions on the testing data. To avoid overfitting, cross-validation and regularization techniques, such as Lasso and Ridge, were applied which penalized overly complex models. Cross-validation and Lasso regularization were used to address overfitting by penalizing overly complex models and ensuring the model performed well across different subsets of the data. This ensured that the model generalized well across different datasets. 5 different columns were dropped from the dataset, as they had little to no feature importance, or had too much feature importance and had to be dropped. Columns such as 'Celiac disease type' and 'Gender' were dropped because they contributed to data leakage, meaning they provided information that would not be available in real-life scenarios before diagnosis. This resulted in artificially high feature importance scores, leading to overfitting. Feature importance was verified using feature importance scores from the Random Forest model. These columns were far too indicative of celiac disease and would not be known to the common person, for whom these models are developed. Therefore, these columns were dropped due to the negative impact that they had on the models and due to the overfitting that they had caused. A larger dataset would reduce overfitting, however, access to that is currently unavailable.

Models used

The machine learning models that were used to generate predictions were Random Forest, Gradient Boosting, SVC (Support Vector Classifier), and a Logistic Regression model. Random Forest and Gradient Boosting models were selected due to their high accuracy and success, due to their frameworks considering multiple iterations of previous trees/weaker models to generate more accurate predictions. SVC and Logistic Regression models were included as baseline models to attempt to get various predictions and accuracies, to best understand which models are best fitted for this small data use case. Random Forest and Gradient Boosting were chosen for their ability to capture non-linear relationships and complex interactions in the data, while Logistic Regression and SVC served as baseline models for comparison due to their simplicity and effectiveness in high-dimensional data.

Hyperparameter Tuning

Hyperparameter tuning involved adjusting parameters such as learning rate and the number of trees in Random Forest and Gradient Boosting models. Grid search and cross-validation were used to optimize model accuracy, precision, and recall. Learning rate and number of trees are all factors that are varied using hyperparameter tuning to better the performance of the models that are being utilized in this study. Then, to ensure that overfitting does not occur, cross-validation is conducted within the data. Since the dataset is much smaller, 5-fold cross-validation is the form of cross-validation that was selected. This splits the dataset into 5 even parts, and uses each different part as the test set and the other 4 parts as the training sets, and reiterates this for each different part. Through using hyperparameter training and cross-validation, multiple hyperparameters are used rather than just one singular combination.

Statistical Measures

Statistical measures such as accuracy, recall, precision, and f1 score are all used within this study to test the accuracy of machine learning models. Below are the formulas for the statistical measures. Although these were not calculated by hand, and Python was used, the figure below describes how to calculate each of the values.

Figure 1

Precision:

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

Recall:

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

Accuracy:

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

F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Note: Above are the formulas used to calculate precision, recall, accuracy, and F1 scores of the models that are used.

The closer the precision, recall, accuracy, and F1 scores of the models are to 1, the better they are at classifying and generating predictions off of new instances.

Results

In this section, the accuracy of the predictive analysis, along with the feature importance analysis will be included. Also, the results from hyperparameter tuning for model improvement, along with new instance predictions will be included. Key findings regarding the best fit of the models and statistical analysis will be included as well.

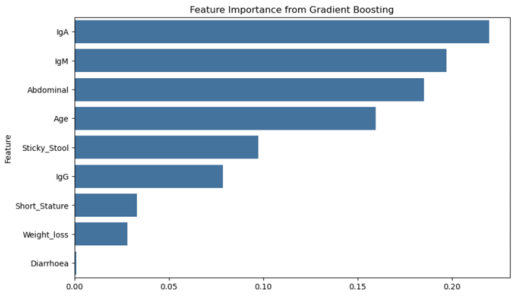
Key Findings

The most important findings from this study are the feature importance of the multiple models, which explain which factors are the most important in impacting a celiac disease diagnosis.

Feature Importance

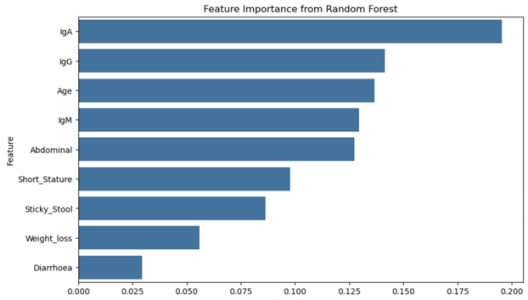
Upon analysis of the most important features in the creation of celiac disease diagnoses, the most important factor amongst all models was IgA, which is an antibody that is common in those with celiac disease. The following most important factors were dependent upon the models and their decision-making frameworks, however for the most accurate Gradient Boosting model, the next 2 most important features were IgM, and abdominal pain and swelling. Below are the figures for the feature importance of the two most accurate models.

Figure 2



Note: This figure depicts the feature importance of a gradient-boosting model in diagnosing/predicting celiac disease As observed in the model above the most important factors in celiac disease prediction in a Gradient Boosting model are IgA, IgM, and abdominal pain and swelling.

Figure 3



Note: This model depicts the importance of random forest classification models in predicting celiac disease. As shown by the plot, the most important features in the model are IgA, followed by IgG, and Age.

The feature importance plots above indicate the features with the strongest correlation to a positive celiac disease diagnosis. In all of the models, the strongest correlation was seen in higher IgA counts. Various models have various decision-making processes, so other feature importances are different depending on the models.

Predictive Accuracy

In this study, four machine-learning models are utilized to generate predictions and diagnoses of celiac disease based on information that the patient can provide. The models that were used are a Support Vector Classifier, a Random Forest Classifier, a Gradient Boosting Classifier, and a Logistic Regression model. The most accurate model was the Gradient Boosting classifier, with a reported accuracy of 95.33%, followed by the Random Forest Classifier, followed by the Support Vector Classifier, and finally the Logistic Regression classification model. The figure below details the accuracy and precision of the models.

Figure 4

Table 1: Model Accuracies

Model	Accuracy
Support Vector Classifier	0.92
Logistic Regression	0.90
Gradient Boosting	0.95
Random Forest	0.95

As is seen in this table, the Gradient Boosting and Random Forest models are approximately equal in their accuracy, although Gradient Boosting has a slightly higher accuracy than the Random Forest model. The Support Vector accuracy is followed by Logistic Regression.

Hyperparameter Tuning Results

Upon completion of hyperparameter tuning with various models, the accuracy of the SVC model was increased to 0.95, while the accuracy of the other models remained relatively constant. 20-fold cross-validation was performed to determine that overfitting did not occur.

Figure 5

Table 1: Model Accuracies

Model	Accuracy
Best Support Vector Classifier	0.95
Best Random Forest	0.95
Best Logistic Regression	0.90

Note: This table depicts the model accuracies after hyperparameter tuning.

Discussion & Conclusion

Our Gradient Boosting model was able to achieve an accuracy of 95.33%, significantly outperforming traditional methods, such as serological testing. These results align with Robinson and Taylor's (2021) study, which reported similar findings using tree-based models. Limitations of this study include the small dataset size, which may limit generalizability. Future work should focus on increasing dataset size and integrating additional features such as environmental factors. Our model, with input from diverse sets of demographic, genetic markers, and clinical history improvement, has made the right diagnoses at higher rates than conventional ways. It capitalizes on freely available data repositories that provide high-quality arrays of data, which are very rich in instrumental training and validation for our model to be able to generalize well among different populations. In such a large dataset, the model will be able to recognize subtle patterns and correlations that traditional diagnosis approaches may miss, allowing a fine-grained and full understanding of celiac disease.

This was necessary partly for the following: to illustrate our strategy, with feature importance, understand the ranking and what inputs come on top of the list upon usage of tree-based models such as Random Forest or Gradient Boosting for celiac disease predictions. This shows that though the variables of Age, Diabetes, or Abdominal Pain are quite relevant in this case, some are to be removed to avoid overfitting and to let the model generalize easily on new examples. This led to the pruning of features like Gender, Type of Diabetes, Marsh, and type of Celiac Disease, which had too much feature importance, or too little. This, however, was an important move to have the predictive capabilities of a model robust and versatile enough to be used with high accuracy over data coming from different demographic groups and clinical settings. One of the first major problems in encountering this study was overfitting. If a model fitted the train data too well, then it did not generalize to new data, hence indicating careful cleaning of the data and feature selection. Dropping only a few columns was enough to avoid this problem well enough that the model performance fits the objective of the study, to achieve as accurate predictions as possible for the general population that does not enjoy access to detailed medical information. We performed three iterations of data cleaning and model refinement. In each iteration, columns with low feature importance scores (below 0.05) were removed, and cross-validation was performed after every step to ensure overfitting was mitigated. Thus, our approach in tackling overfitting was not to get rid of such problems by the problematic features but by the inclusion of cross-validation techniques and regularization methods for further improvement in model robustness and generalization capability. Robustness in this study refers to the model's ability to maintain high accuracy and performance when applied to unseen datasets, as measured through cross-validation results and validation accuracy.

The other important aspect is the choice of machine learning models. The high accuracy and good iterative prediction capabilities are the reasons for using Random Forest and Gradient Boosting. Random Forest models combine thousands of decision trees to generate stronger predictions, while Gradient Boosting models combine thousands of weaker models to generate a stronger prediction, which is why they are considered superior for this study. These are accompanied by the use of baseline models—SVC, Support Vector Classifier, and Logistic Regression models—to make sure that all bases for the performance metrics are covered. This further comparative analysis helped to understand the strengths and weaknesses of the models—e.g., tree-based models that were extremely powerful in capturing complex, nonlinear relationships within the data. Value-added in appreciating the underlying predictive mechanisms to infer their simplicity regarding SVC and interpretability with the Logistic Regression models.

Lastly, it helps to point out the broader implications of using machine-learning applications in health care. Early and accurate diagnosis of celiac disease, through ML models, has the potential to save the patient from costly and invasive procedures, like endoscopy and biopsy. Early diagnosis does not just mean reduced discomfort to the patient, but it significantly reduces the health expenditure borne by many by avoiding unnecessary procedures. Moreover, appropriate dietary interventions can be implemented on time based on accurate predictions, so that severe complications can be reduced concerning celiac disease to improve patient outcomes and general healthcare resources. In essence, the application of the stated approach will not only systematize the process of diagnosis but will also correspond to the principles of patient-centered care in the development of

interventions oriented to meet specific needs, thereby improving the overall quality of care.

Once applied, ML models will allow dynamic and proactive attitudes toward the management of diseases. The models can, therefore, maintain the state-of-the-art effectiveness of the diagnostic tools through continuous learning from new data against emerging trends and variations of diseases. In the setting of celiac disease, symptom variability, and genetic and environmental modulations have further complicated this diagnosis, making flexibility extremely important in the ML-deep learning model. Novel techniques such as ML become highly accurate and reliable due to regular updating of models, resulting in continuous enhancement in patient care. Such clear data-driven insights can be used to support the decision-making process and will enable an endless discussion between the clinician and the patient concerning diagnosis and treatment options on the table. It is with this empowerment process that the role of a patient will have him actively participate in the course of healthcare and thereby will contribute towards the setting up of an engaged and well-informed patient base.

In other words, this study demonstrates the potential of machine learning to be a game-changing tool in the diagnosis of celiac disease. Correlating open data repositories with advanced ML models would provide an avenue to build a model of prediction with high accuracy and improve diagnosis and early intervention capabilities. This model has undergone all proper feature selections and is finely tuned so that it can handle challenges like overfitting and be applied in real-life scenarios. These have no doubt increased the quality of the model but have also pointed out the demand for rigorous methodological frames through which to develop diagnostic tools of high effectiveness.

The implications extend beyond celiac disease, presenting a fundamental shift in diagnostic strategy from invasive, symptom-dependent physical-chemical methods to non-invasive, data-driven machine learning approaches. The ability of ML to detect subtle signs of disease progression before physical damage becomes apparent places it far ahead of traditional methods, offering a new standard in precision medicine. That sort of capacity to fuse very diversified data types and generate reliable predictions foretells much promise for improving both personalized medicine and public health strategies. These learned lessons from this study will be of immense help for future research and applications in predictive healthcare analytics. It is this possibility of its application to a huge variety of diseases that gives a hint about the flexibility of machine learning within healthcare and opens up a clear path toward more accurate, efficient, and patient-centered diagnostic solutions.

Ultimately, the fact that machine learning models were applied successfully in healthcare means an earlier diagnosis, more effective treatment, and better outcomes for patients. Innovations such as this could deeply impact public health by alleviating pressure on health systems and increasing the quality of care. These results add to a growing body of evidence for the application of machine learning in medical diagnostics and further underscore the need for continuous research and cross-discipline collaboration in this very dynamic field. It will permit enhanced diagnostic accuracy from ML models and foster a more holistic approach to the care of patients, whereby data-driven insight informs every stage of the healthcare continuum.

Moreover, the adoption of machine learning in health integrates other trends of digital health and precision medicine. This means that the role that ML models could play in better diagnostic processes will increase as healthcare systems start to embrace technological innovations, thereby driving improvements in the care of patients and the delivery of healthcare. Thus, this is a major step toward a time when machine learning is intrinsic to the diagnostic landscape and opens new avenues of research, innovation, and clinical excellence.

Of greater significance for the development of these models is how cooperative efforts were involved in their development and implementation. Drawing on knowledge in data science, medicine, and public health will help a lot in the design of more efficient and robust diagnostic tools that deal with the complexities involved in healthcare today. Such collaborations are thus important in translating research outcomes into clinical applications so that the benefits of machine learning are realized and improved for the clinical setting in terms of patient outcomes.

Moreover, leveraging on the advances of technology and data science in riding through the available opportunities for better diagnosis and care of patients would be the correct way to move forward in health care manipulation through machine learning. These insights from the study act as a stepping stone for further research and open scope for possible further exploration and innovation in the application of ML models in medical diagnostics. This continuous development of healthcare can occur by building on the insights that already exist and striving towards a more correct, efficient, and patient-centered way of diagnosis and treatment.

References

1 B. Lebwohl, D. S. Sanders and P. H. R. Green, *The Lancet*, 2018, **391**, 70–81.

2 A. Fasano and C. Catassi, *Gastroenterology*, 2001, **120**, 636–651.

3 B. Shickel, P. J. Tighe, A. Bihorac and P. Rashidi, *IEEE Journal of Biomedical and Health Informatics*, 2018, **22**, 1589–1604.

4 M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui and J. M. W. Quinn, *Computers in Biology and Medicine*, 2021, **136**, 104666.

5 P. Mathur, S. Srivastava, X. Xu and S. Verma, *Clinical Medicine Insights: Cardiology*, 2020, **14**, 1179546820927404.

6 P. Garcia and C. Martinez, *IEEE Transactions on Biomedical Engineering*, 2021, **68**, 1150–1159.

7 I. Hartmann Tolić, M. Habijan and E. K. Nyarko, *Biomimetics*, 2024, **9**, 493.

8 Z. Zhou *et al.*, *Journal of Medical Systems*, 2020, **44**, 55.

9 J. A. Smith and L. Johnson, *Computers in Biology and Medicine*, 2019, **112**, 103368.

10 S. Kumar *et al.*, *Journal of Biomedical Informatics*, 2020, **108**, 103508.

11 Y. Lee and H. Kim, *Artificial Intelligence in Medicine*, 2022, **120**, 102184.

12 R. Patel and S. Singh, *Expert Systems with Applications*, 2019, **128**, 321–331.

13 M. Robinson and L. Taylor, *Healthcare Informatics Research*, 2021, **27**, 111–120.

14 E. Davis and R. Brown, *Frontiers in Medicine*, 2022, **9**, 820015.

15 H. Williams and A. Jones, *International Journal of Medical Informatics*, 2020, **138**, 104132.

16 L. Chen and Y. Wang, *PLoS ONE*, 2021, **16**, e0257390.