Explainable Classification of Patients with Primary Hyperparathyroidism Using Highly Imbalanced Clinical Data Derived from Imaging and Biochemical Procedures

Agorastos-Dimitrios Samaras 1, Maria Tsimara 2, Sofia Voidila 2, Nikolaos Papandrianos 1, Petros Zampakis 2, Serafeim Moustakidis 1,3, Elpiniki Papageorgiou 1,* and Christina Kalogeropoulou 2

1 Department of Energy Systems, University of Thessaly, Gaiopolis Campus, 41500 Larisa, Greece; agsamaras@uth.gr (A.-D.S.); npapandrianos@uth.gr (N.P.); smoustakidis@gmail.com (S.M.)
2 Department of Radiology, University of Patras, 26504 Patras, Greece; tsimalykes@yahoo.gr (M.T.); sofiavoidila5@gmail.com (S.V.); pzampakis@upatras.gr (P.Z.); ckrat@upatras.gr (C.K.)
3 AIDEAS OÜ, 10117 Tallinn, Estonia
* Correspondence: elpinikipapageorgiou@uth.gr

Abstract: Primary hyperparathyroidism (PHPT) is a common endocrine disorder characterized by hypercalcemia and elevated parathyroid hormone (PTH) levels. The most common cause is a single parathyroid adenoma, though the rest of the cases are due to multiglandular disease [double adenoma/hyperplasia]. The main focus driving this work is to develop a computer-aided classification model relying on clinical data to classify PHPT instances and, at the same time, offer explainability for the classification process. A highly imbalanced dataset was created using biometric and clinical data from 134 patients (six total features, 20.2% multiglandular instances). The features used by the current study are age, sex, max diameter index, number of deficiencies, Wisconsin index, and the reference variable indicating the type of PHPT. State-of-the-art machine learning (ML) classification algorithms were used in order to create trained prediction models and give predicted classifications based on all features/indexes. Of the ML models considered (Support Vector Machines, CatBoost, LightGBM, and AdaBoost), LightGBM was able to procure the best performing prediction model. Given the highly imbalanced nature of the particular dataset, oversampling was opted for, so as to increase prediction robustness for both classes. The ML model’s performance was then evaluated using common metrics and stratified ten-fold validation. The significance of this work is rooted in two axes: firstly, in the incorporation of oversampling to smooth out the highly imbalanced dataset and offer good prediction accuracy for both classes, and secondly, in offering an explainability aspect to an otherwise black-box ML prediction model. The maximum achievable accuracy for adenoma is 86.9% and for multigland disease 81.5%. Summarizing the above, this study demonstrates the potential for an ML approach to improve the diagnosis of PHPT and also highlights the importance of explainable artificial intelligence (AI).

Keywords: explainability; primary hyperparathyroidism; machine learning; computer-aided diagnosis

1. Introduction

In modern society, PHPT is a quite common health disorder. Its primary characteristic is the excessive secretion of parathyroid hormone (PTH) by one or more of the parathyroid glands [1]. This results in hypercalcemia, bone resorption, and renal dysfunction [2]. The most common cause of PHPT is a single parathyroid adenoma, accounting for approximately 80% of cases [3]. Other causes of PHPT may also include hyperplasia of the
parathyroid glands, multiple adenomas, and parathyroid carcinoma. The diagnosis of PHPT is confirmed by elevated serum calcium and PTH levels. The treatment of choice is surgical removal of the affected parathyroid gland(s) [4]. In recent years, there has been a trend towards less invasive approaches, such as minimally invasive parathyroidectomy and focused parathyroidectomy. However, careful preoperative localization is necessary to achieve high success rates with these techniques [5].

Accepted preoperative localization utilizes ultrasound and sestamibi scans, which have reported sensitivities for abnormal parathyroid localization of up to 74% and 58%, respectively [6]. More recently, dynamic multiphase computed tomography (4DCT) using four phases (non-contrast and post-contrast arterial, early and delayed venous phases) shows higher sensitivities of 55–87%. Furthermore, two- or three-phase scans have shown similar results [7]. In a meta-analysis, the overall pooled sensitivity of CT for localization of the pathological parathyroid(s) to the correct quadrant was 73% (95% CI: 69–78%), which increased to 81% (95% CI: 75–87%) for lateralization to the correct side [8].

On the other hand, recent breakthroughs in the field of AI have made it possible to utilize AI tools in healthcare, revolutionizing the field by providing more accurate diagnoses, predicting diseases, and improving patient outcomes. Customized Medical Decision Support Systems (MDSS) are a prime example of how AI is being used in healthcare [9]. These systems employ machine learning algorithms to analyze patient data and provide personalized diagnoses and treatment options. With the ability to analyze large amounts of data, such systems have the potential to improve accuracy in the detection of diseases, allowing for earlier intervention and treatment. Additionally, AI can assist in the development of new drugs, by predicting the behavior of molecules and identifying potential targets for therapy. With the use of AI, physicians can spend more time with patients, leading to better patient experiences and outcomes. While AI in healthcare is still in its early stages, it has the potential to transform the industry and improve patient care, creating a more efficient and effective healthcare system.

The subject of PHPT prediction using ML models has been tackled by the research community, mainly with the use of image data [10]. One such study [11] employed a deep learning algorithm to classify PHPT based on clinical, laboratory, and imaging data. The study found that a modified DenseNet architecture was able to accurately classify PHPT with a sensitivity of 90.6% and a specificity of 94.5%. In ref. [12], a boosted tree classifier, RandomTree, was used to predict PHPT based on clinical and laboratory data, with a maximum accuracy of 94.1%. Moreover, researchers in [13] used ML models to predict the risk of postoperative hypocalcemia in patients undergoing parathyroidectomy for PHPT. This study found that the prediction model was able to accurately differentiate an abnormal versus a normal parathyroid gland with a precision–recall characteristic curve of 0.93. Finally, another study [14] used deep learning (DL) methodologies, in order to develop a prediction model for persistent PHPT after parathyroidectomy. The study found that the prediction model had a maximum achievable sensitivity of 96.3% and a specificity of 97%, suggesting that it could be a useful tool for guiding postoperative management. Overall, these studies demonstrate the potential for ML to improve the diagnosis and management of PHPT. However, further research is needed to validate these findings and develop ML-based tools that can be effectively integrated into clinical practice.

It immediately becomes clear that in the scientific pool of relevant work, the vast majority of studies use datasets consisting of, or accompanied by, image data. In contrast, the current work utilizes a small feature set, that focuses mainly on three indexes: the maximum diameter of the deficiency index, the number of deficiencies index, and the Wisconsin index [15]. These three essential indices play pivotal roles in the evaluation and characterization of non-small cell lung cancer (NSCLC) cases. The maximum diameter index serves as a crucial metric, representing the largest dimension of a given lesion or tumor. It is instrumental in quantifying the size and extent of pathological findings, contributing to a comprehensive assessment of the lesions under consideration. The number of deficiencies index provides a quantitative measure of the identified abnormalities, offering
insights into the count and distribution of deficiencies within the studied cohort. This index aids in understanding the overall burden and diversity of pathological manifestations, enriching the dataset analysis. The reported deficiencies are referring to the number of abnormal disorders of parathyroid glands depicted in ref. [16]. This means that when the number of deficiencies is greater than one they are defined as multigland diseases [17,18]. Additionally, the Wisconsin index, hailing from a well-established lineage of indices, encapsulates a comprehensive evaluation of NSCLC cases. Derived from a combination of features and measurements, this index reflects a different perspective on the complexities of the disease, thus contributing valuable dimensions to the overall dataset analysis. Together, these indices collectively enhance the study’s statistical analyses and contribute to a more subtle understanding of NSCLC characteristics.

We believe that by utilizing a small feature set, a highly adaptive but accurate methodology for PHPT classification can be created. Moreover, due to the highly imbalanced nature of PHPT instances, general accuracy may be deceiving. A trained model could predict, very accurately, just the single tumor class and not the other classes. However, as this class consists of over 80% of the entries in most cases, the overall accuracy will be high, obscuring, hence, the fact that the model could not accurately predict the rest of the classes. To achieve high accuracy for all classes, we incorporated stratified oversampling so as to train the ML model more efficiently for the classification process. The prediction results are evaluated using common metrics and stratified 10-fold cross-validation [19]. Finally, in the scope of the current study, we explain the prediction process of the most accurate classification model, and we discuss these findings in tandem with the medical literature on the matter of PHPT. Therefore, all in all, the three major contributions of our work can be summed up as follows: (1) the use of only clinical data and indexes to classify PHPT instances; (2) implementing oversampling on the dataset to overcome its imbalanced nature; and, finally, (3) exploring and explaining the black-box mechanics of an ML algorithm to be used in a computer-aided decision-making system for PHPT diagnosis and classification.

The organization of the rest of this paper goes as follows: Section 2 outlines the details of the patient dataset, the ML algorithms employed for classification, and the features used for the prediction procedure. The evaluation process used to measure the performance of the models is also outlined in this section, along with the proposed explainability analysis. In Section 3, the experimental results are presented, including the performance metrics and any notable findings. Section 4 delves deeper into the strengths and weaknesses of the proposed model, offering an explanation of the best-performing model and finding common patterns with the established medical bibliography on the matter of PHPT. Finally, Section 5 concludes this study by summarizing the key findings and suggesting potential avenues for future research.

2. Materials and Methods

2.1. Patient Population

The dataset that was used in the current work involved 134 participants. Of this pool of subjects, 27 have multiglandular parathyroid disease (20.15% class MG), and the rest are affected by adenoma. Additionally, 16.42% of the patients are male, and as far as the age factor is concerned, there was a range of 23 to 84 years of age. Finally, this study opts for a minimalist approach on the matter of features, as apart from the afore-mentioned demographic data, it includes just three more medical indexes. These are, namely, the maximum diameter index, the number of deficiencies index, and the Wisconsin index. The entirety of the features is presented in Table 1.
Table 1. Features used as input by the prediction model, after binary normalization.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
<th>Feature Class/Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>Male/female</td>
</tr>
<tr>
<td>2</td>
<td>&lt;40</td>
<td>Aged under 40</td>
</tr>
<tr>
<td>3</td>
<td>40–50</td>
<td>Aged between 40 and 50</td>
</tr>
<tr>
<td>4</td>
<td>51–60</td>
<td>Aged between 51 and 60</td>
</tr>
<tr>
<td>5</td>
<td>&gt;60</td>
<td>Aged over 60</td>
</tr>
<tr>
<td>6</td>
<td>Max diameter index</td>
<td>Maximum diameter of findings</td>
</tr>
<tr>
<td>7</td>
<td>Num of defs</td>
<td>Number of findings</td>
</tr>
<tr>
<td>8</td>
<td>Wisconsin index</td>
<td>Wisconsin index</td>
</tr>
<tr>
<td>9</td>
<td>Parathyroid gland</td>
<td>Adenoma or multiglandular (MG)</td>
</tr>
</tbody>
</table>

The patients comprising this work’s dataset were previously biochemically diagnosed with primary hyperparathyroidism and subsequently underwent multi-detector computed tomography (MDCT) for the presurgical localization of abnormal parathyroid glands in the period of 2016–2021.

The MDCT protocol is a 2-phase protocol involving a pre- and post-contrast injection phase, the latest at 40 s (late arterial phase), covering the area of the neck and the superior mediastinum. Images were analyzed on a workstation (vitreos), and abnormal parathyroid glands were recorded according to their location, number, and size of the glands.

Multiglandular disease was recorded as (2) when more than one abnormal gland was identified, though a single abnormal gland was identified as (1). Furthermore, according to the maximum diameter of the abnormal parathyroid gland, a categorization of 0 (>13 mm), 1 (7–13 mm), and 2 (<7 mm) was also applied, namely, the maximum diameter index.

The Wisconsin index created by Mazeh et al. [15] on the other hand, is derived from the multiplication of the PTH value with the blood calcium value and categorized as 0 (>1600), 1 (800–1600), and 2 (<800) [17]. All patients were treated surgically, and findings were based on the number of abnormal parathyroid glands that were identified and classified as [S] when a single gland was removed and as [MG] in the case of multiglandular disease.

Data collection has been approved by the ethical committee of the University General Hospital of Patras (Ethical and Research Committee of University Hospital of Patras). The retrospective nature of this study waives the requirement to obtain informed consent from the participants. All data-related processes were performed anonymously. All procedures in this study were in accordance with the Declaration of Helsinki.

2.2. Features

The dataset used in the context of the current study is a minimalistic one. Being based on five features and one reference feature, this dataset exclusively employs clinical data, without any image data whatsoever. This adds to the versatility and applicability of the prediction model, as no special information is needed for a possible application of the classification mechanism to most relative datasets.

On the matter of age, we opted for a form of normalization. Given that the rest of the fields had integer values, we split up the age information into four different fields. More specifically, the 4 different fields used for age are <40, 40–50, 50–60, and >60. This way, it is also easier to highlight common patterns between specific age ranges and each class of this scenario (adenoma/MG).

The most important information for this dataset is included in the three medical indexes used: the max diameter index, the number of deficiencies, and the Wisconsin index. All the above indices have been proposed in ref. [17] as part of a scoring system for the prediction of multigland disease in primary hyperparathyroidism. According to their study, the bigger the score (corresponding to the summation of all the above indices), the higher the probability of multigland disease.

In particular, as reported in the literature, the imaging method of four-dimensional computed tomography (4D-CT) was used to build a scoring model based on the features
of 4D-CT, including largest lesion size and the number of suspicious lesions. The 4D-CT MGD score achieved a good specificity of 81–96% and variable sensitivity of 39–64% for predicting MGD. Moreover, the previous study reported that patients with multigland disease had significantly lower Wisconsin index scores, smaller lesion size, and a higher likelihood of having either multiple or zero lesions identified on 4D-CT.

2.3. ML Algorithms

For the purposes of the current study, four different classification models, each based on a different ML algorithm, were tested. Support Vector Machine (SVM), Categorical Boosting machine (CatBoost), Light Gradient-Boosting machine (LightGBM), and Adaptive Boosting machine (AdaBoost) were the well-documented ML algorithms that were employed in the scope of this work. This range of algorithms has been widely used by the research community [20–27] with proven effectiveness in numerous medical scenarios.

The SVM classifier [28] is a popular algorithm in machine learning used for classification and regression analysis. SVM is a supervised learning model that identifies the optimal hyperplane to separate two classes of data with the maximum margin. This approach makes SVM particularly useful for complex datasets. The algorithm has been used in various domains, including finance, healthcare, and natural language processing. Particularly, SVM has been used to diagnose breast cancer, to predict stock market trends, and to classify sentiment in social media data.

CatBoost [29] is a gradient boosting algorithm that has become increasingly popular in recent years. The algorithm is designed to handle categorical data and can automatically preprocess this type of data, making it particularly useful for natural language processing tasks. CatBoost has been used in various domains, including healthcare, finance, and e-commerce. Various CatBoost applications include, among others, predicting breast cancer prognosis, identifying fraudulent credit card transactions, and recommending products to online shoppers.

On the other hand, LightGBM [30] is a gradient boosting framework that utilizes decision tree algorithms to improve the speed and accuracy of machine learning models. It was developed by Microsoft researchers in 2017, and has since gained popularity among data scientists and researchers due to its speed and accuracy. The algorithm uses a unique approach to build trees by splitting the data points that contribute the most to the loss function, which results in a more efficient and accurate model. LightGBM has been used in various applications, including image recognition, natural language processing, and financial forecasting.

Adaptive Boosting [31], or as it is commonly known, AdaBoost, is a classification ML algorithm designed by Yoav Freund and Robert Schapire in 1995. It is a popular ensemble learning algorithm that combines multiple weak learners to create a stronger model. AdaBoost assigns weights to each training instance, and then trains weak learners to classify the data. The algorithm then combines the weak learners to create a final model. AdaBoost has been used in various domains, including computer vision, finance, and healthcare. For example, AdaBoost has been used to detect faces in images, to predict stock prices, and to diagnose heart disease.

2.4. Oversampling

Oversampling is a technique used in AI to address the quite common issue of imbalanced datasets. This occurs in classification scenarios, where one class of data is significantly underrepresented compared to the rest. It involves increasing the number of instances of the minority class by randomly generating synthetic samples. Oversampling can help to improve the performance of machine learning models, particularly in scenarios where the minority class is of high importance, such as in medical diagnosis. There are several common oversampling methods, including Random Oversampling [32], the Synthetic Minority Oversampling Technique (SMOTE) [33], and Adaptive Synthetic Sampling (ADASYN) [34]. Studies have shown that oversampling techniques can significantly improve the accuracy
of machine learning models in imbalanced datasets [33], demonstrating the importance of considering this technique in the development of AI models.

All of the abovementioned oversampling techniques were considered and tested with the particular dataset. Nevertheless, random oversampling gave the most balanced accuracy. In the current classification scenario, all the employed prediction models gave the most equally accurate predictions for both classes (adenoma/MG), when they were trained on datasets modified by random oversampling. Since a prediction model can easily be trained to be accurate for the prevalent class, generating a model that is equally accurate for both classes is of utmost importance. The holistic approach of this study, along with all its steps, is showcased in Figure 1.

Random oversampling [32] is one of the most straightforward and widely used oversampling methods. It involves randomly duplicating instances of the minority class until it reaches a certain threshold, thus balancing the classes. Although it has limitations and is prone to overfitting, random oversampling is a simple and effective way to address imbalanced datasets. The method was first introduced by Chawla et al. in 2002 [33] and has been widely used in various domains, including finance, healthcare, and cybersecurity. For example, random oversampling has been used to improve the accuracy of fraud detection models in credit card transactions [35], to predict early signs of Alzheimer’s disease [36], and to predict depression among subjects [37].

Figure 1. Current study’s methodology.
2.5. Evaluation of Results

The next stage of this study is to evaluate the prediction results of each prediction model in order to eventually discern the optimal one. To this end, six widely implemented metric scores were used: accuracy [38], sensitivity [39], specificity [39], Jaccard score [40], F1-score [41], and confusion matrix [42]. All of the above metrics determine the performance of a model by using a combination of correlation between true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances.

Traditionally, the most important metric in prediction studies is considered to be accuracy. Still, this cannot be the case in this particular dataset. This being a highly imbalanced dataset, basing the selection of the optimal model mainly on accuracy can be deceiving [43–45]. Due to the fact that one class is considerably more prevalent, a prediction model could be trained to predict the prevalent class very well, while the other(s) are not predicted well. This shortcoming, nevertheless, is not accurately depicted in the overall accuracy of the model. Therefore, when dealing with imbalanced datasets, the most important features tend to be the confusion matrix, sensitivity/specificity for each class, and the receiver operating characteristic (ROC) curve [46]. It is possible that oversampling techniques may result in a lesser overall accuracy for the above reasons. They do, however, create a more all-around trained estimator that is eventually more suitable to classify every instance instead of just the prevalent class.

Specificity and sensitivity are two important measures that help evaluate the performance of a diagnostic test. Sensitivity is the proportion of true positives that are correctly identified by the test, while specificity is the proportion of true negatives that are correctly identified by the test. These measures are important when evaluating the effectiveness of a medical test or a decision-making tool. This can be especially true for classification scenarios of two classes in total, such as the current one (adenoma/MG classification). These metrics can provide insight into the model’s ability to correctly identify individuals who belong to the first class (sensitivity; in the current scenario this is for adenoma) and those who do not (specificity). The mathematical formulas used to calculate the aforementioned metrics are expressed as follows:

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \]  
\[ \text{Specificity} = \frac{TN}{TN + FP} \]

On the other hand, the ROC curve is a useful graphical tool employed to evaluate the performance of a binary classification model. It plots the true positive rate (sensitivity) against the false positive rate (specificity) for various threshold values, allowing the user to select a threshold that balances the trade-off between sensitivity and specificity. The area under the ROC curve (AUC) is a commonly used metric to quantify the overall performance of a classification model. A perfect model will have an AUC of 1, while a random model will have an AUC of 0.5 [47].

Lastly, a confusion matrix is a table that summarizes the performance of a classification model by comparing predicted and actual outcomes. In binary classification problems, such as in this work, it is a 2 by 2 matrix that is formed by the TP, TN, FP, and FN fields. Such a confusion matrix has the following form:

<table>
<thead>
<tr>
<th>Actual condition</th>
<th>Positive (P)</th>
<th>Negative (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (PP)</td>
<td>True positive (TP)</td>
<td>False negative (FN)</td>
</tr>
<tr>
<td>Negative (N)</td>
<td>False positive (FP)</td>
<td>True negative (TN)</td>
</tr>
</tbody>
</table>

In order to further validate the performance of each tested prediction model, a 10-fold stratified cross validation was employed. Hence, these cross-validated results were used to
calculate the performance of each ML model. Eventually, the optimal prediction model was selected not on merit of best overall accuracy, but on merit of the above analyzed metrics. To evaluate the data distribution’s normality, the Shapiro–Wilk test was applied. Categorical variables underwent chi-square analysis, while quantitative variables were analyzed using an independent samples t-test, with a confidence interval set at 95%. This analysis facilitated a thorough assessment of both qualitative and quantitative variable aspects, thereby establishing a statistical foundation for identifying the features most critical to the classification of adenoma versus multiglandular conditions.

2.6. Interpretation of Results

The final step in the methodology of this study is the explanation of the optimal prediction model. This analysis is of critical importance for all black-box prediction models, as it offers the users of such systems valuable insight into the decision-making process of the AI. This is especially true with AI systems that are being deployed in medical applications. Specifically, in the healthcare field, ML systems that lack explainability may not be perceived as trustworthy by patients or medical personnel [48–50] and they may fail to comply with certain rules and regulations, which have explanations as a mandatory requirement for automated decision-making systems [51]. Therefore, if AI is to be considered as a valuable tool for health experts, it is of the outmost importance it and its prediction process be less opaque.

This study takes on the matter of explaining the chosen black-box ML prediction model, by utilizing two main analytical approaches in tandem: Cohen’s effect sizes and SHAP analysis. These are two widely established techniques that are being used in a variety of scientific fields to analyze datasets and explain the results of experiments [52–55]. Moreover, the explained results can be compared to established bibliographies and knowledge in the field of the application and offer further validation to the process of prediction/classification of the particular computer-aided decision-making tool.

Cohen’s effect size is a statistical metric that indicates the magnitude of the difference between two groups. It is calculated by dividing the difference between the means of the groups by the pooled standard deviation. Cohen (1988) [56] suggested that an effect size of 0.2 should be considered small, 0.5 medium, and 0.8 large. This measure is widely used in different fields, including psychology, medicine, and other sciences, to interpret the results of experiments and studies. It provides a valuable tool for assessing the practical significance of the findings and comparing them to previous research.

On the other hand, SHAP analysis [57] is a widely used technique that aims to increase the interpretability and transparency of machine learning models. This approach is based on cooperative game theory concepts and treats each feature as a “player” in a game where the prediction is the goal. By measuring the impact of each feature on the prediction outcome, SHAP analysis assigns an importance value to each feature, which allows for the identification of the most relevant features in the model. This feature-level importance assessment provides a valuable tool for understanding and interpreting the model’s predictions and can help increase user trust in the model. Furthermore, SHAP analysis can be used to detect potential biases and help ensure that ML models are fair and reliable.

3. Results

This study, as described in the Methods section employed the steps depicted in Figure 1 in order to analyze data in a Linux environment, specifically an 8-core i7 CPU, 16GB DDR3 RAM, and Ubuntu 20.04LTS. This project utilized the Python programming language as its core coding framework, with several machine learning-specific libraries, including sklearn, shap, and genetic_selection, proving particularly useful in creating an automated environment capable of evaluating the performance of multiple machine learning prediction models. By utilizing the capabilities of these libraries, the study
was able to take full advantage of their machine learning features to produce robust and reliable results.

3.1. ML Models’ Performance

Table 2 showcases the specificity, sensitivity, and accuracy metrics achieved by each ML model when using 10-fold cross validation on the dataset as it is (with no oversampling whatsoever). Overall, the prediction model based on LightGBM can be considered to have performed the best. It achieved the best results in terms of specificity and overall accuracy. Specifically, it predicted multiglandular cases with an accuracy of 66.67% (tied for best with the CatBoost based model) and achieved an accuracy of 83.58%. On the other hand, the SVM based model was the most accurate at predicting adenoma cases, with a sensitivity of 91.59%.

Table 2. Stratified CV-10 metrics for each ML model—without oversampling.

<table>
<thead>
<tr>
<th>Metric</th>
<th>SVM</th>
<th>CatBoost</th>
<th>LightGBM</th>
<th>AdaBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>91.59%</td>
<td>86.91%</td>
<td>87.85%</td>
<td>90.65%</td>
</tr>
<tr>
<td>Specificity</td>
<td>44.44%</td>
<td>66.67%</td>
<td>66.67%</td>
<td>40.74%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>82.09%</td>
<td>82.84%</td>
<td>83.58%</td>
<td>80.60%</td>
</tr>
</tbody>
</table>

Table 3 showcases the specificity, sensitivity, and accuracy metrics achieved by each ML model. Overall, the prediction model based on LightGBM is considered to be the optimal one for the particular dataset. It achieved the best results on all metrics compared to the other models. Specifically, it predicted single gland (adenoma) cases accurately in 86.92% and multiglandular cases in 81.48%, and achieved an overall accuracy (after being trained on an oversampled dataset) of 85.82%. Additionally, Figure 2 depicts the confusion matrices for each prediction model.

Table 3. Stratified CV-10 metrics for each ML model—after oversampling.

<table>
<thead>
<tr>
<th>Metric</th>
<th>SVM</th>
<th>CatBoost</th>
<th>LightGBM</th>
<th>AdaBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>85.05%</td>
<td>86.92%</td>
<td>86.92%</td>
<td>84.11%</td>
</tr>
<tr>
<td>Specificity</td>
<td>77.78%</td>
<td>77.78%</td>
<td>81.48%</td>
<td>77.78%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>83.58%</td>
<td>85.07%</td>
<td>85.82%</td>
<td>82.84%</td>
</tr>
</tbody>
</table>

Furthermore, the Shapiro–Wilk test confirmed that all features conform to a normal distribution. Our statistical analysis revealed that the only feature of statistical significance was the number of deficiencies ($p = 0.000$). Importantly, the other parameters did not demonstrate significance, with $p$-values exceeding 0.05, indicating their limited statistical impact on distinguishing between adenoma and multiglandular conditions.

3.2. Explainability Plots for the Optimal Prediction Model

The optimal model for this set of experiments was the one based on the LightGBM algorithm. It achieved an overall accuracy score of 85.82%, a sensitivity score of 86.92%, and a specificity score of 81.48%. Figure 3 depicts Cohen’s effect sizes for each feature, and Figure 4 displays the relative summary plot using the SHAP values for every feature used as input.
Furthermore, the Shapiro–Wilk test confirmed that all features conform to a normal distribution. Our statistical analysis revealed that the only feature of statistical significance was the number of deficiencies ($p = 0.000$). Importantly, the other parameters did not demonstrate significance, with $p$-values exceeding 0.05, indicating their limited statistical impact on distinguishing between adenoma and multiglandular conditions.

3.2. Explainability Plots for the Optimal Prediction Model

The optimal model for this set of experiments was the one based on the LightGBM algorithm. It achieved an overall accuracy score of 85.82%, a sensitivity score of 86.92%, and a specificity score of 81.48%. Figure 3 depicts Cohen's effect sizes for each feature, and Figure 4 displays the relative summary plot using the SHAP values for every feature used as input.

In Figure 3, the features are ranked based on their impact on the prediction result. The higher the impact, the higher the feature appears in the plot. More specifically, the number of deficiencies index feature has the highest impact on the prediction result, followed by the Wisconsin index and the max diameter index. On the other hand, the sex of the subject is characterized by the lowest overall impact.

In a similar tone, in the summary plot of SHAP values (Figure 4), the features are also sorted in order of magnitude of effect. Positive SHAP values indicate that the particular range of values for a specific feature leads the prediction towards the MG class. For instance, when the value of the feature for the number of deficiencies is high, it contributes towards an MG classification. Instead, when the maximum diameter index has high values, it contributes towards an S classification (i.e., adenoma).

Figure 2. Confusion matrices of the ML prediction models; (a) depicts SVM, (b) depicts Catboost, (c) depicts LightGBM, and (d) depicts AdaBoost.

Figure 3. Cohen's effect size for each input feature sorted by magnitude of effect.
4. Discussion

The driving objective behind this work is multifaceted. First of all, an accurate computer-aided decision-making system was to be generated that could predict all the classes of the particular imbalanced dataset equally accurately. Tables 2 and 3 and Figure 2 highlight how, after being trained on an oversampled set, each prediction model demonstrated a more balanced prediction accuracy for both classes. Furthermore, the next goal of this work was to shed light on the decision-making process of such a black-box computer-aided decision-making model. On this scope, Figures 3 and 4 offer valuable insight as to which features were the most essential for the classification and how each datum contributed to the final outcome. Below, we will try to analyze the findings more analytically, explain what these plots showcase, and finally comment on the findings from a medical viewpoint.

The comparison of evaluation metrics (Tables 2 and 3) for the two implementation scenarios of this study, with oversampling during the training and without, highlights an increase in the robustness of the models. Indeed, every model showcases better all-around prediction performance when the random oversampling technique is used during its training (Table 3). More specifically, without oversampling the sensitivity of every model was higher than when employing oversampling (Table 2). This, basically, means that the prediction systems were better trained to predict adenoma cases. However, there is a significant increase in the specificity with oversampling and an increase in the overall accuracy of the models as well. Therefore, with oversampling, every model could not only definitely classify MG cases much more accurately but also showcase an overall enhancement in performance. The reason behind this is straightforward, since the oversampled training
subset includes more instances of the least prevalent class, it allowed the ML algorithm to be trained in a more balanced way.

On the other hand, the comparison between the features identified as most influential by Cohen’s effect sizes and their corresponding SHAP values reveals a high degree of similarity (Figures 3 and 4, respectively). The results suggest that the number of deficiencies is the most impactful feature, followed by the Wisconsin index and the maximum diameter of the deficiency. These findings are consistent with the known risk factors for PHPT, as documented in the medical literature [58]. Overall, the Cohen’s effect size chart and the SHAP summary plot provide valuable insights into the factors influencing PHPT and are consistent with existing medical knowledge.

The diagnosis of MGD preoperatively is quite difficult, though it is very substantial to reduce the risk of surgery failure. Even with the use of all available imaging/clinical data, it is quite hard to predict the presence of multiple parathyroid lesions in some cases. Previous studies established scoring systems based on imaging/clinical data for predicting MGD. Sepahdari et al. [17] built a scoring model based on the features of 4D-CT, in combination with biochemical information, achieving a specificity of 81–96% and a variable sensitivity of 39–64% for predicting MGD. More recently, Yanwen et al. [18] developed a nomogram based on US findings and clinical factors to predict MGD in PHPT patients, with a specificity of 0.94 and a sensitivity of 0.50.

The number of abnormal glands identified in imaging studies is the most important factor indicating MGD. In our study, the number of abnormal glands (feature “number of defs”) proved to have the best predilection toward MGD. The Wisconsin index shows a predilection towards MGD, as is expected from the existing literature [15], since a low Wisconsin index is indicative of MGD. High values of the max diameter index (indicating small glandular size) surprisingly show a predilection toward SD, instead of MGD, as expected. The reason for this paradoxical event is two-fold. On the one hand, MGD can be the result of double or even triple large-sized adenomas. Moreover, SD can comprise a small-sized adenoma. Therefore, a high-value max diameter index (i.e., size of glands) did not prove to be a reliable prognostic factor for MGD, as there is really no clear cut-off value regarding gland size, to distinguish MGD from SD [16]. Sex and age do not show any clear predilection toward MGD. This is in accordance with the previous literature [58].

To sum up, incorporating explainability techniques such as Cohen’s effect sizes and SHAP values is crucial for gaining a better understanding of the model’s decision-making process and the factors that affect its predictions. This is particularly important in order to receive maximum trust from the medical community, as they need to have confidence in the accuracy of the model’s predictions. Furthermore, the insights gained from these techniques can be used to identify outlier cases and potential errors in input data, leading to more accurate predictions and improved diagnosis of PHPT. Overall, incorporating explainability mechanisms into machine learning models is essential for making them more reliable and useful in practical applications.

The results also underscore the advanced capabilities of our proposed explainable ML methodology, distinguishing it from conventional statistical approaches. Traditional statistical analysis might overlook certain features, deeming them insignificant. However, our methodology reveals that these features, when integrated with other statistically significant parameters, play a crucial role in the ML model’s decision-making process. Notably, both statistical and explainability analyses in this paper concurred on the importance of the number of deficiencies. Yet, the explainability analysis further unveils that the Wisconsin index, age over 60, and maximum diameter index also significantly influence the model’s outcomes, despite not being identified as statistically significant. This highlights the added value of explainability analysis in identifying critical contributing factors that traditional methods might neglect, offering deeper insights into the complex interactions that drive the ML model’s predictions.

Nevertheless, this study has certain limitations. Particularly, the data used as input in the current work lack image data. Image data hold crucial information that is widely
used by medical experts to assess the health condition of a patient and the type of PHPT. Hence, it is highly likely that incorporating such images into the input data, would further improve the prediction accuracy of a decision model for PHPT classification. Furthermore, the number of features used in this study is small. On the one hand, this makes the current application more adaptable, but on the other hand, it offers limited information on the patients. It is possible that an extended feature set could lead to even more accurate results.

Concluding, in the context of healthcare and clinical practices, it is crucial to acknowledge that ML-based tools are still in their infancy. While these tools exhibit promising capabilities in various domains, they should be viewed as supportive aids rather than replacements for the expertise of medical professionals. It is imperative to emphasize that ultimate medical decisions should be entrusted to qualified healthcare practitioners who possess the necessary clinical knowledge and experience. The proposed model, presented in this study, serves as a valuable and helpful tool that can assist medical experts in their decision-making processes. However, it is essential to exercise caution and responsibility when integrating such technologies into sensitive healthcare matters. The collaboration between machine learning models and medical professionals can potentially enhance diagnostic processes.

5. Conclusions

The main objective of this work was to overcome the restrictions a highly imbalanced dataset presents for classification and to create an all-around accurate classification ML system for PHPT instances. Indeed, a clear increase in specificity was observed, ranging from 11.11% to 37.04%, and the overall accuracy was enhanced as well. Additionally, much-needed insight into a black-box prediction system for PHPT classification was derived. The incorporation of explainability mechanisms such as SHAP analysis and Cohen’s effect sizes provides transparency into the decision-making process of the model. This allows for a better understanding of the factors that contribute to its predictions and helps establish trust between the AI model and the human medical user. By highlighting the features that have the most significant impact on the prediction, the user can assess the reasoning behind the prediction and potentially identify any errors or outliers in the input data. This is particularly crucial in the medical field, where the accuracy and reliability of predictions can have a significant impact on patient outcomes. Therefore, incorporating explainability mechanisms can enhance the usefulness and trustworthiness of AI models in medical decision making. Another key takeaway of this study is that the AI classification system is consistent with contemporary research on PHPT, as evidenced by the alignment of the most common discerning factors for PHPT, with high SHAP values and Cohen’s Effect sizes. This can enhance confidence in the system, particularly when it produces unexpected outputs. Moreover, the explainability outcomes provide legitimacy to the computer-aided decision-making system as they are consistent with years of research findings. Moving forward, our future plans include incorporating image data into our prediction models or incorporating prediction results from AI models that utilize image data as input. Moreover, we intend to conduct further research on the findings of Section 2, particularly on additional well-established PHPT risk factors.


Funding: This research received no external funding.

Institutional Review Board Statement: Data collection has been approved by the ethical committee of the University General Hospital of Patras (Ethical and Research Committee of University Hospital of Patras). Approval Code: 11391/27-4-18.
Informed Consent Statement: The retrospective nature of the study waives the requirement to obtain informed consent from the participants. All data-related processes were performed anonymously. All procedures in this study were in accordance with the Declaration of Helsinki.

Data Availability Statement: The data presented in this study are available in the article.

Conflicts of Interest: Author Serafeim Moustakidis was employed by the company AIDEAS OÜ. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.