Primary Hyperparathyroidism Screening with Machine Learning

An Honors Thesis(HONR 499)

by

Noah Ziems

Thesis Advisor

Dr. Shaoen Wu

Ball State University
Muncie, Indiana

May 2021

Expected Date of Graduation

May 2021

Primary Hyperparathyroidism Screening with Machine Learning

Noah Ziems

Abstract

Primary Hyperparathyroidism(PHPT) is a relatively common disease, affecting about one in every 1,000 adults. However, screening for PHPT can be difficult, meaning it often goes undiagnosed for long periods of time. While looking at specific blood test results independently can help indicate whether a patient has PHPT, often these blood result levels can all be within their respective normal ranges despite the patient having PHPT. Based on clinical data from the real world, in this work, we propose a novel approach to screening PHPT with neural network (NN) architectures, achieving over 97% accuracy with common blood values as inputs. Further, we propose a second model achieving over 99% accuracy with additional lab test values as inputs. Moreover, compared to traditional PHPT screening methods, our NN can reduce the false negatives of traditional screening methods by 99%.

Acknowledgments

Process Analysis Statement

To the passing observer, this may seem like a random thesis. Why does a Computer Science student care so much about screening for an tumor?

In March, 2018 I was in the middle of my second semester at Ball State. Right after Spring Break ended I began feeling sick. Unfortunately this sick feeling didn't go away and caused me to miss lots of class. To make matters worse, anytime I went to the doctor they could not find anything wrong with me. I clearly appeared to be having some sort of heath issue but no one could pin down exactly what was going on. As my illness continued, I decided to take the fall semester off to focus on my health. However, even after taking a semester off we still were not able to figure out what was wrong, so I was forced to go back to school and do the best I could. My illness continued for another year until my doctor finally sent me to Mayo Clinic in Rochester, Minnesota. I ended up spending about a month at Mayo Clinic doing many tests every week, with each test showing up as normal. They couldn't find anything wrong, so I returned back to Indiana having missed an entire month of school and having nothing to show for it.

A month later I got an unexpected call from a doctor at Mayo Clinic saying that he was reviewing my chart and found some abnormalities. He wanted to see me in Rochester as soon as possible. After looking at my test results, he had noticed some abnormal blood results that none of the other doctors had seen. These abnormal blood results went all the way back to March 2018, almost two years earlier, when I first began having health issues. My second trip to Mayo was again filled with tests, but this time each test result came back abnormal. We had found what was wrong. It turns out a tumor had developed on a parathyroid gland in my neck that was causing me to produce too much of a particular hormone called parathyroid hormone. This hormone was causing an overabundance of calcium to leech from my bones into my bloodstream, causing issues with every cell in my body and making me feel sick. The surgery was very easy and within a few weeks I was back to normal.

After going through any health issue like that, any sane person would want to leave that part of their life behind them. The completely rational thing to do is to move on, and for some time I did that. However, after a while I began obsessively investigating this disease, trying to figure out why it took so long for my doctors to figure out what was wrong with me. After digging more into why this disease is notoriously difficult to diagnose, I had an idea. As a machine learning researcher, I was used to automating detection of things that are hard to detect. At the time I was working on a project funded by Lockheed Martin to detect security vulnerabilities in code. I could easily build a similar statistical model that predicts whether someone had the disease from some very common blood lab values. Due to its algorithmic nature, this model would likely be better and faster at identifying patients with Primary Hyperparathyroidism (PHPT) than any doctor. This would be my way of

getting back at this disease. It would be revenge for taking away two years of my life. The big uncertainty was not debating whether this was possible or not, but instead finding a surgery center that would be willing to work with an undergraduate ML researcher with no proven track record.

Luckily, I found Dr. Jim Norman who runs the Norman Parathyroid Center(NPC). NPC does more parathyroid surgeries for treating PHPT than any other surgery center in the world. I sent Dr. Norman a cold email telling him my story and asking if I could use his data to make a model that would screen for PHPT. I was sure he would not reply to my email so I began looking into other alternatives like Mayo Clinic or some other big hospital system. To my surprise, I got a response a day later from Dr. Norman with more enthusiasm about the project than I myself had at the time. Within a week he had sent me all his data from more than 30,000 patients that had received their parathyroid surgeries at NPC. With this data I had everything I needed to make the model.

The results of the research are best shown in the paper below. Aside from getting the data, the biggest challenge of this research was investigating the model's behavior in odd circumstances. For example, our model focuses on parathyroid hormone(PTH) levels between 20 and 100mg/dl. However, there are times when a patient with parathyroid cancer may have a PTH level of over 200. In other words, there are other diseases that can contribute to a high calcium and PTH level, which are not accounted for in our model. We tried to find data for parathyroid cancer, but it's such a rare occurrence that even if we were able to find it, we wouldn't be able to train a model that could correctly identify it. Other diseases could also impact CA and PTH levels including kidney failure and vitamin d deficiency. To solve this, we emphasize the model is intended for screening and not diagnosis. The model is intended to alert a doctor when they should consider further testing, and is not meant to be a replacement for a doctor's diagnosis.

I think the biggest thing I learned from this research is the importance of taking your time when doing research in healthcare. This is even more important at the intersection of Machine Learning and healthcare. When doing work in healthcare alone, there are huge implications when something does not work as intended, so one has to be very careful to make sure it always behaves as expected. Machine learning models are notorious for being unpredictable when presented with data outside the training distribution. They also tend to "fail silently," meaning you can't always tell when they are doing something wrong. Therefore, when combining machine learning with healthcare, one has to intentionally look for any odd behaviors that may cause serious issues when the model is used in practice.

Introduction

Primary Hyperparathyroidism (PHPT) is a disease caused when a parathyroid develops a tumor causing the parathyroid to over produce parathyroid hormone (PTH), which regulates the amount of calcium in the bloodstream. Too much calcium in the bloodstream for extended periods of time leads to various health issues such as osteoporosis, kidney dysfunction, high blood pressure, and others. Moreover, PHPT is relatively common, affecting about one in every 1,000 adults. However, the current medical practice of manual screening for PHPT is very difficult. While identifying whether a patient has PHPT can be facilitated by independently inspecting a few variables of their blood lab results, often these blood lab values can all be within normal ranges despite a patient having the disease. A more reliable way to screen for PHPT, therefore, is to consider the values of the blood test variables coherently as opposed to independently. Because these blood values in lab results are not often significant individually, millions of patients with PHPT have been missed in their traditional annual physical exams that have blood lab results. Only when they show PHPT symptoms later will their lab results will be examined by a specialist that can diagnose PHPT. Therefore, it is important to design an automated AI solution to diagnose the patients with PHPT from their blood lab results in regular visits or physical exams much earlier than symptoms begin to show.

Deep learning with neural networks(NNs) has been shown to be surprisingly effective at learning relationships among data, often surpassing human performance in data-driven tasks such as object detection and recognition (Russakovsky et al. 2015). Deep learning driven health informatics in many areas such as cancer diagnosis has achieved amazing progress in recent years with many solutions reportedly outperforming human medical experts (Esteva et al. 2021). In this work, we present a novel machine learning based solution for screening PHPT in patients using simple neural networks. This work is a collaboration between computer science and a world-class PHPT clinic with the real medical lab result data of thousands of patients.

More specifically, our solution demonstrates the following advantages over traditional PHPT screening methods:

1. **Speed**: By using automated algorithmic methods that can be run on a computer, each individual patient is evaluated in a matter of milliseconds where a trained doctor may take minutes or more to interpret in traditional screening methods.

2. **Cost**: The computational cost of evaluating each patient is significantly less than $0.01, while using a doctor with traditional screening methods costs significantly more.

3. **Scale**: Instead of evaluating patients one by one as a doctor would, our model is able to run in parallel, evaluating many patients at once.

4. **Accuracy**: Our method proves itself to be highly accurate while allowing for future improvements with more data. However, it remains to be seen how this accuracy compares to that of an experienced medical expert.

5. **Precision**: Due to the deterministic nature of computer software, the same results are always given for the same inputs. Where doctors may differ in opinion, our method will always produce the same diagnosis given the same inputs.

Related Work

2.1 Machine Learning in Health Informatics

Significant progress has been made using state-of-the-art deep learning based computer vision techniques in health informatics (Bellamy, Celi, and Beam 2020). Wu et al. develop a computer vision model, finding it as accurate as experienced radiologists in screening breast cancer (Wu et al. 2019). Moreover, there are a number of benchmark datasets that have shown impressive improvements in performance for disease diagnosis, including CheXpert (Irvin et al. 2019), SD-198(Sun et al. 2016), and the International Skin Imaging Collaboration (ISIC) dataset (Rotemberg et al. 2020).

However, relatively little progress has been made on more traditional tabular data that most medical data is comprised of, typically in the form of Electronic Health Records (EHRs). There is some debate as to why machine learning has not had as much impact in health informatics as in other fields. New studies show this is likely caused by a lack of benchmarks in the field, preventing machine learning researchers from comparing new methods to traditional ones (Bellamy, Celi, and Beam 2020).

2.2 PHPT Screening in Endocrinology

If there is no prior suspicion of PHPT, screening of PHPT is only done when the patient has a metabolic panel drawn to check his or her general health. A high serum calcium level from a metabolic panel is often an indication of PHPT. However, very often doctors disregard this because many other factors can contribute to high calcium levels. The true normal range of serum calcium varies based on other factors, such as age and gender, which are not often taken into account by traditional screening methods. Instead, these traditional screening methods only evaluate a patient based on the normal range of a blood lab value for the general population irrespective of the other blood lab values. How to accurately diagnose PHPT in an integrative consideration of all these factors remains open, which also motivates our work of achieving such a goal with machine learning.

Clinical Dataset

In this work, we propose using machine learning to accurately diagnose PHPT by considering the factors including serum calcium level that can be obtained in blood lab results, age and gender. The data supporting our machine learning model consists of two separate parts. The first part comes from real PHPT patient data while the second part is synthetic data generated using the observed distribution metrics of patients without PHPT.

The data stemming from real patients comes from the Norman Parathyroid Center (Norman, Goodman, and Politz 2010), which annually conducts the most parathyroid surgeries in the United States. The dataset is comprised of the data from 20,000 patients having positive diagnosis of PHPT. In the data, each record consists of parameters including age, gender, pre-operative serum calcium level, and pre-operative parathyroid (PTH) hormone, all of which are the input to our machine learning PHPT screening model, as well as some other information about each patient. The extra information is not taken into account by our model because their measurements are only taken after a patient is suspected to have PHPT and thus would not fall into the category of screening. It is worth noting that each patient in the dataset had the serum calcium and parathyroid hormone levels measured at least 3 times which are then averaged before they are put into the dataset. During training time, the record of each patient in this portion of the dataset is labeled as having PHPT. In the dataset, patient privacy information such as name and address has been stripped off.

To supplement our dataset with non-PHPT data, which is necessary for training a machine learning model, we have generated synthetic data based on the previously observed distributions of serum calcium and parathyroid hormone in patients without PHPT. Each synthetic data point is given a calcium level drawn with a Gaussian distribution of $\mu = 9.6$mg/dl and $\sigma = 0.15$ and a PTH level with a Gaussian distribution of $\mu = 34$pg/ml and $\sigma = 4.5$ (Norman, Goodman, and Politz 2010). To maintain a dataset that is independent and identically distributed(IID), the synthetic data has the exactly same number of data points as the real patient data, yielding a total dataset of around 40,000 data points.

Based on the distributions our synthetic data is drawn from, half of the data labeled as PHPT-Negative is male with the other half being female. For those who are labeled in the dataset as PHPT-Positive, 76% are female. This may appear to indicate a dataset which is not independently and identically distributed (IID). However, this proportion is in line with previously observed distribution of PHPT Positive incidence in clinical settings (MW et al. 2013).

Figure 3.1 shows the distributions of the important continuous data used by our models. Areas in blue indicate PHPT-Positive distributions where areas in or-

ange indicate PHPT-Negative distributions. The distributions have relatively little overlap, which further indicates machine learning is a promising solution for PHPT screening. It is worth noting that the distribution of PHPT-Negative in Figure 3.1a is intentionally uniform from the assumption that age of the general population is uniform as opposed to Gaussian.
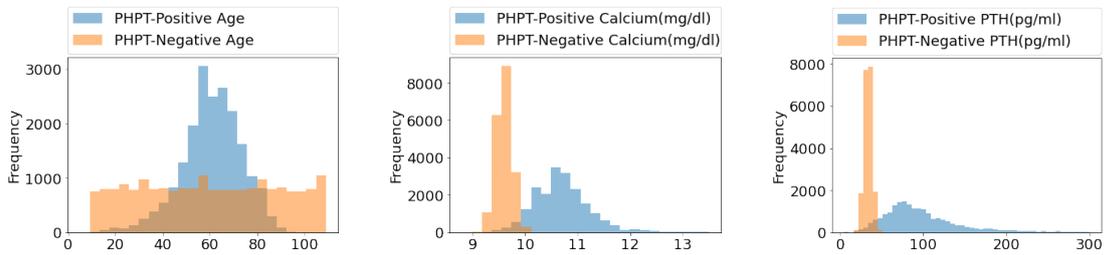
(a) The distributions of age in our dataset. Note the PHPT-Positive distribution is from real patients with PHPT where the PHPT-Negative distribution is generated from a uniform distribution.

(b) The distributions of calcium in our dataset. Note the PHPT-Positive distribution is from real patients with PHPT where the PHPT-Negative distribution is generated from a normal distribution.

(c) The distributions of PTH in our dataset. Note the PHPT-Positive distribution is from real patients with PHPT where the PHPT-Negative distribution is generated from a normal distribution.

Figure 3.1: The distributions of age, calcium and PTH in the clinic dataset

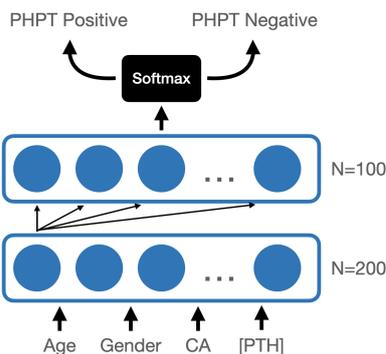# Neural Network Screening for Primary Hyperparathyroidism

Figure 4.1: The generic neural network architecture used by both models

This paper proposes using neural network (NN) models to automate the PHPT screening based on four factors: age, gender, pre-operative serum calcium level, and pre-operative parathyroid (PTH) hormone level. Due to the relatively low dimensionality of the input data and output data, a simple neural network architecture has been chosen. We have designed and implemented two nearly-identical NN models for PHPT screening, with the only difference being the number of inputs. The generic architecture of the PHPT NN is illustrated on Figure 4.1.

4.1   PHPT Neural Network Inputs

For both models, gender, age, and average serum calcium(Avg Ca) are taken as inputs. However, one of the models also takes average parathyroid hormone(PTH) into account while the other does not. To avoid any confusion, we refer to these models as *PTH-Included* and *PTH-Ignored*. For both models, gender is treated as a categorical variable that is one-hot encoded. Age, average calcium, and average PTH are all treated as continuous inputs.

The motivation of designing two different identical NN models lies in medical application use cases. The goal of this work is to improve screening for PHPT. PTH is a critical component in diagnosing PHPT, however it is rarely tested if there is no prior suspicion of PHPT. In contrast, serum calcium is far more commonly tested in regular blood lab work. Therefore, we create another model, *PTH-Ignored*, which only takes inputs that are common in regular lab tests for patients who are not suspected to have PHPT. The *PTH-Ignored* model can then be used to screen a

much broader set of patients in hopes of finding more PHPT patients even if there is no prior suspicion of PHPT.

4.2   PHPT Neural Network Hidden Layers

As stated above, both NN model architectures are identical in their internal structures. They both have two hidden layers with the first layer containing 200 neurons and the second layer containing 100 neurons. A ReLU(Nair and Hinton 2010) is used in between these two layers, serving as the necessary nonlinear activation function.

4.3   PHPT Neural Network Outputs

The output of the last hidden layer is a vector that is run through a Softmax function to find the final classification probabilities as the final output of the NN for PHPT screening. For both models, there are only two output classes, one for PHPT-Positive and the other for PHPT-Negative.

4.4   PHPT Neural Network Model Training

For the PHPT NN model training, the dataset is split using standard 5-fold cross validation. The training set accounts for 80% of the original dataset and the test set accounts for the remaining 20%. For each training iteration, the patient lab values in the training dataset are given as input to the models. After the models computes the outputs, a loss function is used to measure the deviation of the prediction from its real value. The loss then backpropagates through the models and weights are updated so the models are more correct next time they are given similar data as inputs. After many thousands of iterations, the models converge on a particular set of weights that has minimal loss. In each of our experiments, the models are trained for five epochs with a triangular cyclical learning rate and minimal hyperparameter tuning. It is worth noting that both models are trained in the exactly same way with the only difference being on the input of the PTH values or not.

Performance Evaluation
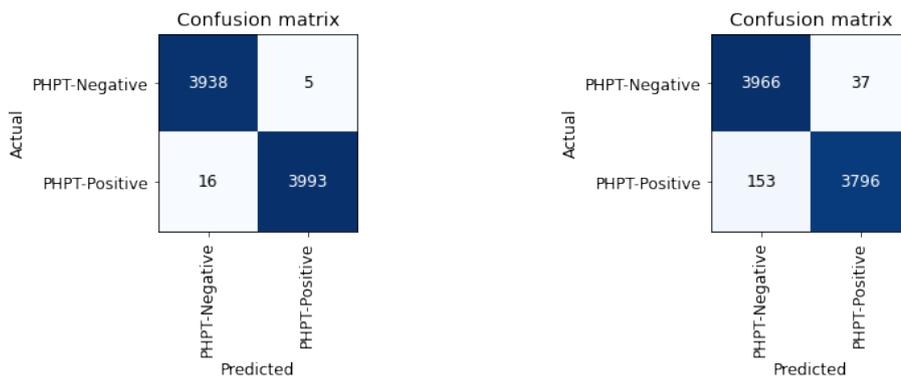
(a) A confusion matrix showing our model's predictions versus the ground truth of the PHT-Included model

(b) A confusion matrix showing our model's predictions versus the ground truth of the PHT-Ignored model

Figure 5.1: Confusion Matrices

5.1 Experiment Platform and Settings

We have implemented and evaluated our PHPT NN models with PyTorch (Paszke et al. 2019). On the top of PyTorch we use FastAI (Howard et al. 2018) due to it's existing API designed for tabular data, which evaluating our dataset requires. After training the models on the training dataset, we then assessed them on the test dataset, which is data the trained models have never seen before. This test dataset is meant to represent the data the models would come across when used in practice.

For fair comparison, we have evaluated the same dataset using traditional screening methods. For traditional PHPT screening, any calcium level between 8.4 mg/dl and 10.5 mg/dl is considered normal and anything outside of that is flagged as abnormal. For PTH, anything between 9 pg/ml and 69 pg/ml is considered normal while anything outside that range is considered abnormal.

5.2 Performance Metrics

The key performance metrics used to evaluate our models are *accuracy, precision*, and *recall. Accuracy* in our case indicates how often our models' prediction is correct

in recognizing a patient has PHPT based on the input clinic data. Higher accuracy is better. *Precision* indicates the percentage of positive predictions our models correctly predict a patient has PHPT based the input data. *Precision* is much more sensitive to false positives than accuracy. *Recall* is very similar to *Precision*, but is instead sensitive to false negatives. The definition and calculation for *accuracy*, *precision*, and *recall* are shown below.

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

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

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

Here TP stands for true positive, TN for true negative, FP for false positive and FN for false negative.

5.3   Confusion Matrices

The confusion matrices in Figure 5.1 show all data needed to compute the results shown in Table 5.1. The numbers in the top left and bottom right in dark blue are TP and TN indicating how many datapoints are correctly classified in the dataset whereas the numbers in the top right and bottom left in light grey are FP and FN indicating how many datapoints in the dataset are misclassified. As shown in Figure 5.1a, the PTH-Included model has significantly fewer false negatives when compared to the PTH-Ignored model shown in Figure 5.1b.

5.4   Results and Observations

The results of our experiments are summarized in Table 5.1.

Architecture	Accuracy	Precision	Recall
PTH-Included	99.73%	99.87%	99.60%
PTH-Ignored	97.61%	99.03%	96.12%
Traditional Screening	93.06%	100%	88.69%

Table 5.1: Performance summary of PHPT screening with two NN models and traditional screening

From the results, it can be observed that the screening accuracy of both models far surpassed the performance of traditional screening methods. It is worth noting that the PTH-Included model performed significantly better than the PTH-Ignored model, as the PTH-Included model has access to more information relevant to diagnosis than the PTH-Ignored model. This is an intentional design choice.

A surprising observation is that, although the traditional screening method had 0 false positives after evaluating all patients in the dataset, it had over 2,000 false negatives. In other words, a significant portion of patients with PHPT in the dataset would not have been recognized as having PHPT. In contrast, the PHPT-Included

NN model only had 5 false positives and 16 false negatives. Therefore, our best PHPT NN model can reduce the false negatives of traditional expert-screening by 99.27%.

The accuracy of both PHPT NN models suggest that screening for PHPT with machine learning is not only possible, but could be incredibly effective in practice. Moreover, the total marginal cost of evaluating a single patient is only from the cost of electricity to support the computation of the model that takes on average seven milliseconds on our current hardware.

Model Capabilities and Limitations

The diagrams on Figure 6.1 are prediction maps showing how the model's predictions change when given different inputs. These can also be thought of as the decision boundaries, where yellow indicates high confidence that a patient has PHPT and blue indicates high confidence that a patient does not have PHPT. Green areas indicate where the model is unsure whether the patient has PHPT or not based on the data given.

6.1 Model Capabilities

Shown in Figure 6.1a and Figure 6.1b, our model begins flagging patients as PHPT positive when calcium goes above 10.5 mg/dl or when PTH goes above 55pg/ml. However, the decision boundary changes shape as the input age increases, flagging patients as PHPT Positive with lower calcium and PTH levels. This is inline with the observed PHPT incidence rate as age increases, particularly with post-menopausal women(MW et al. 2013). Traditional screening methods do not often take this into account, ignoring age as a factor all together.

Moreover, Figure 6.1d and Figure 6.1e show the PTH-Ignored model's prediction map when comparing male and female patients. The decision boundary is slightly lower with females than males, which is also in line with observed distributions in clinical data. In both cases, the CA level needed to flag a patient as PHTP positive decreases with age, with the decrease being slightly more aggressive for females around age 40.

It is worth noting all of these decision boundaries are statistically learned with no input from the authors aside from the generated synthetic data.

6.2 Model Limitations

We have further investigated the limitations and potential issues of the PHPT NN models when the training data is out of expected distribution. As shown in Figure 6.1c, the models can yield mistakes when exposed to data that is substantially out of the distribution it has been trained on. For age 60 and older, the probability of PHPT begins to decrease even when average serum calcium levels increase. To our best knowledge, this is medically incorrect and likely caused by the uniform distribution that was used for age in the synthetic portion of our dataset. Further, when the average serum calcium goes below 9.0 as in Figure 6.1d, the model begins increasingly classifying patients as PHPT-Positive. This also is likely incorrect.
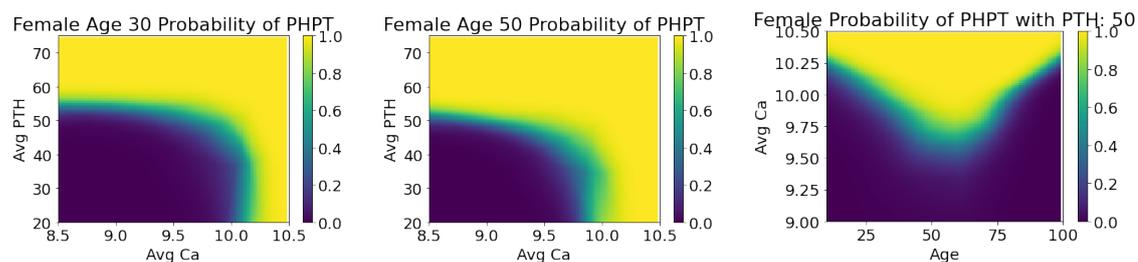
(a) A prediction map of PHPT given a female age 30. Yellow indicates high probability of PHPT while purple indicates low probability.

(b) A prediction map of PHPT given a female age 50. Yellow indicates high probability of PHPT while purple indicates low probability.

(c) A prediction map of PHPT given a female with PTH 50. Yellow indicates high probability of PHPT while purple indicates low probability.
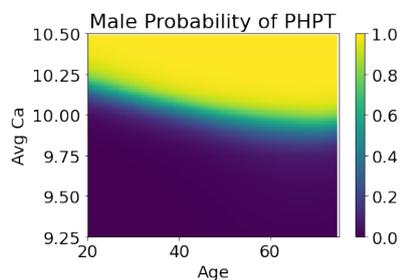
(d) A prediction map of PHPT for a male. Note this is for the PTH-Ignored model.
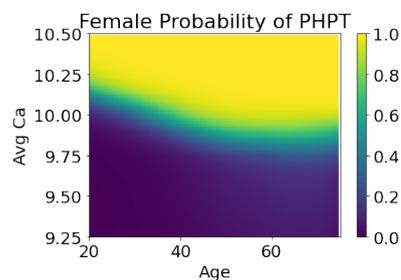
(e) A prediction map of PHPT for a female. Note this is for the PTH-Ignored model.

Figure 6.1: Model Predication Maps

Both of these problems would be solved with real data sourced from patients that do not have PHPT, rather than using the synthetic data.

Conclusions

This paper proposes an automated PHPT screening solution with neural network machine learning models. With the real clinic data, the solution shows the ability of simple neural networks to achieve surprising results when trained to screen for PHPT. Compared to traditional screening, our solution can significantly improve the accuracy while it can reduce the false negative by 99%. Moreover, we show our approach to screening for PHPT has numerous advantages to the traditional methods, including speed, scale, and consistency. The robustness and reliability of the model can be even more improved with real data of patients without PHPT.

Bibliography

Russakovsky, Olga et al. (2015). *ImageNet Large Scale Visual Recognition Challenge.* arXiv: `1409.0575 [cs.CV]`.

Esteva, Andre et al. (2021). "Deep learning-enabled medical computer vision". In: *Digital Medicine.*

Bellamy, David, Leo Celi, and Andrew L. Beam (2020). *Evaluating Progress on Machine Learning for Longitudinal Electronic Healthcare Data.* arXiv: `2010.01149 [cs.LG]`.

Wu, Nan et al. (2019). *Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening.* arXiv: `1903.08297 [cs.LG]`.

Irvin, Jeremy et al. (2019). *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.* arXiv: `1901.07031 [cs.CV]`.

Sun, Xiaoxiao et al. (2016). "A benchmark for automatic visual classification of clinical skin disease images". In: *European Conference on Computer Vision.* Springer, pp. 206–222.

Rotemberg, Veronica et al. (2020). "A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context". In: *arXiv preprint arXiv:2008.07360.*

Norman, J, A Goodman, and D Politz (2010). "Calcium, parathyroid hormone, and vitamin D in patients with primary hyperparathyroidism: normograms developed from 10,000 cases." In: *Endocrine practice : official journal of the American College of Endocrinology and the American Association of Clinical Endocrinologists.*

MW, Yeh et al. (2013). "Incidence and prevalence of primary hyperparathyroidism in a racially mixed population." In: *The Journal of Clinical Endocrinology and Metabolism.*

Nair, Vinod and Geoffrey E. Hinton (2010). "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning.* ICML'10. Haifa, Israel: Omnipress, pp. 807–814. ISBN: 9781605589077.

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32.* Ed. by H. Wallach et al. Curran Associates, Inc., pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Howard, Jeremy et al. (2018). *fastai.* `https://github.com/fastai/fastai`.