Automated Primary Hyperparathyroidism Screening with Neural Networks

Noah Ziems
Department of Computer Science
Ball State University
Muncie, IN USA
nmziems@bsu.edu

Shaoen Wu
School of Information Technology
Illinois State University
Normal, IL USA
swu1235@ilstu.edu

Jim Norman
Norman Parathyroid Center
Tampa, FL USA
jimnormanmd@gmail.com

***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 the clinic data from the real world, in this work, we propose a novel approach to screening PHPT with neural network (NN) architecture, 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 models can reduce the false negatives of traditional screening methods by 99%.**

I. Introduction

Primary Hyperparathyroidism (PHPT) is a widely existing 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 test variable values can all be within normal ranges despite a patient having PHPT. 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 variable 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 of utmost importance 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 [10]. 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 [2]. 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.

II. Related Work

A. Machine Learning in Health Informatics

Significant progress has been made using state-of-the-art deep learning based computer vision techniques in health informatics [1]. Wu et al. develop a computer vision model, finding it as accurate as experienced radiologists in screening breast cancer [12]. Moreover, there are a number of benchmark datasets that have shown impressive improvements in

performance for disease diagnosis, including CheXpert [4], SD-198[11], and the International Skin Imaging Collaboration (ISIC) dataset [9].

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 [1].

B. 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.

III. 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 [7], 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 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 learned 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 $\sigma = 4.5$ [7]. 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 [5].

Figure 1 shows the distributions of the important continuous data used by our models. Areas in blue indicate PHPT-Positive distributions where areas in orange 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 1a is intentionally uniform from the assumption that age of the general population is uniform as opposed to Gaussian.

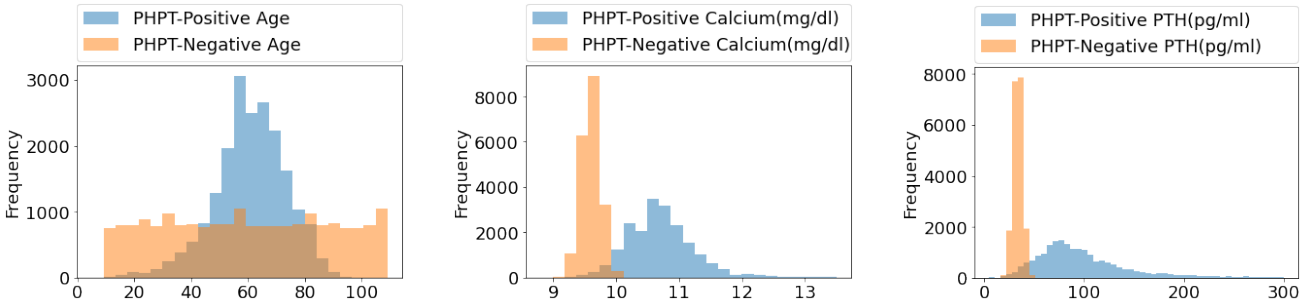
IV. Neural Network Screening for Primary Hyperparathyroidism

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 2.

A. 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



(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.

Fig. 1: The distributions of age, calcium and PTH in the clinic dataset

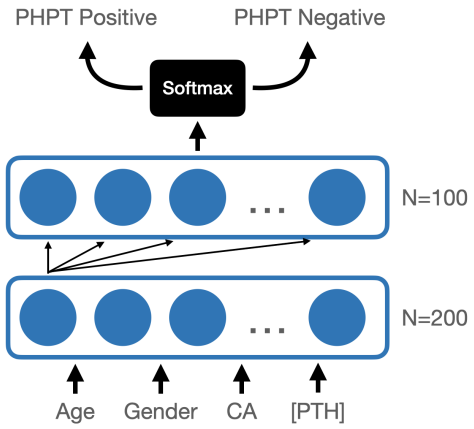


Fig. 2: The generic neural network architecture used by both models

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.

B. 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[6] is used in between these two layers, serving as the necessary nonlinear activation function.

C. 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.

D. PHPT Neural Network Model Training

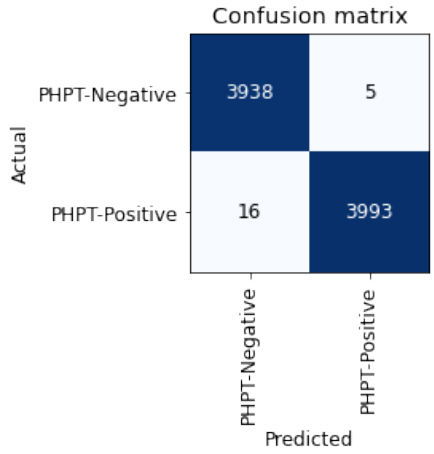
For the PHPT NN model training, the dataset is split using standard 5-fold cross validation. The training set accounts for 80% of the original dataset and the test set accounts for the remaining 20%. For each training iteration, the patient lab values in the training dataset are given as input to the 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.

V. Performance Evaluation

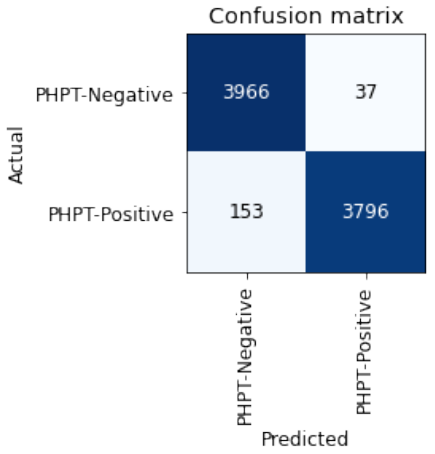
A. Experiment Platform and Settings

We have implemented and evaluated our PHPT NN models with PyTorch [8]. On the top of PyTorch we use FastAI [3] 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



(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

Fig. 3: Confusion Matrices

pg/ml is considered normal while anything outside that range is considered abnormal.

B. 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.

C. Confusion Matrices

The confusion matrices in Figure 3 show all data needed to compute the results shown in Table I. 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 3a, the PTH-Included model has significantly fewer false negatives when compared to the PTH-Ignored model shown in Figure 3b.

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

D. Results and Observations

The results of our experiments are summarized in Table I.

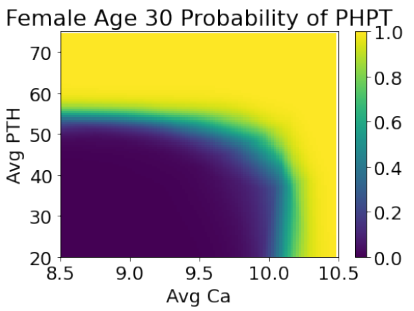
From the results, it can be observed that the screening accuracy of both models far surpassed the performance of traditional screening methods. It is worth noting that the PTH-Included model performed significantly better than the PTH-Ignored model, as the PTH-Included model has access to more information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

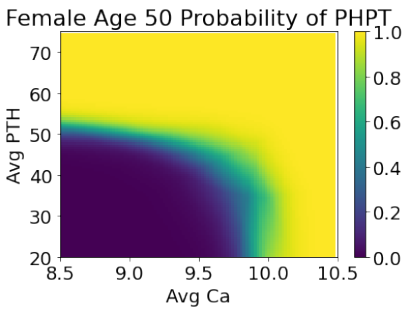
The accuracy of both PHPT NN models suggest that screening for PHPT with machine learning is not only possible, but could be incredibly effective in practice. Moreover, the total marginal cost of evaluating a single patient is only from the cost of electricity to support the computation of the model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

VI. Model Capabilities and Limitations

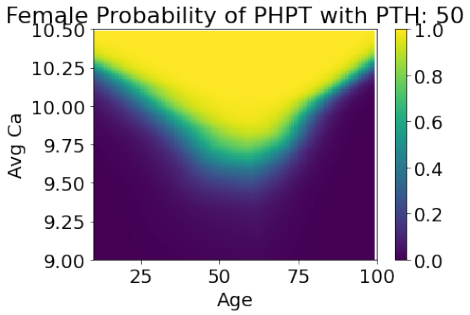
The diagrams on Figure 4 are prediction maps showing how the model's predictions change when given different inputs.



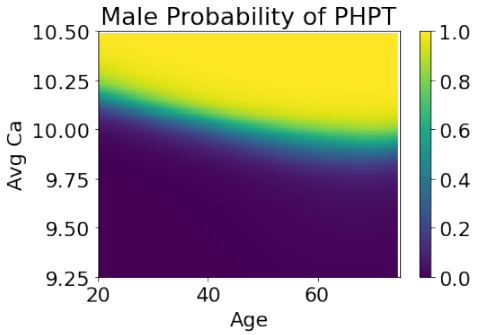
(a) A prediction map of PHPT given a female age 30. Yellow indicates high probability of PHPT while purple indicates low probability.



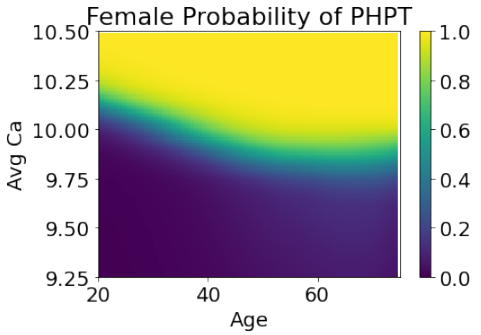
(b) A prediction map of PHPT given a female age 50. Yellow indicates high probability of PHPT while purple indicates low probability.



(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.

Fig. 4: Model Predication Maps

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.

A. Model Capabilities

Shown in Figure 4a and Figure 4b, 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[5]. Traditional screening methods do not often take this into account, ignoring age as a factor all together.

Moreover, Figure 4d and Figure 4e 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.

B. 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 4c, 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 4d, the model begins increasingly classifying patients as PHPT-Positive. This also is likely incorrect. Both of these problems would be solved with real data sourced from patients that do not have PHPT, rather than using the synthetic data.

VII. 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 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.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. #2109971.

REFERENCES

[1] D. Bellamy, L. Celi, and A. L. Beam. Evaluating progress on machine learning for longitudinal electronic healthcare data, 2020.

[2] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher. Deep learning-enabled medical computer vision. *Digital Medicine*, 2021.

[3] J. Howard et al. fastai. https://github.com/fastai/fastai, 2018.

[4] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

[5] Y. MW, I. PH, Z. HC, N. S, L. IL, H. A, H. PI, and A. AL. Incidence and prevalence of primary hyperparathyroidism in a racially mixed population. *The Journal of Clinical Endocrinology and Metabolism*, 2013.

[6] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress.

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

[8] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[9] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *arXiv preprint arXiv:2008.07360*, 2020.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

[11] X. Sun, J. Yang, M. Sun, and K. Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016.

[12] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzębski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras. Deep neural networks improve radiologists' performance in breast cancer screening, 2019.