DermaChat Development

Miguel Silva
Masters in Artificial Intelligence Engineering (MEIA)
Polytechnic of Porto – School of Engineering (ISEP)
Porto, Portugal
1201045@isep.ipp.pt

Hugo Simão
Masters in Artificial Intelligence Engineering (MEIA)
Polytechnic of Porto – School of Engineering (ISEP)
Porto, Portugal
1222686@isep.ipp.pt

Vitor Silva
Masters in Artificial Intelligence Engineering (MEIA)
Polytechnic of Porto – School of Engineering (ISEP)
Porto, Portugal
1030578@isep.ipp.pt

Mariana Ribeiro
Masters in Artificial Intelligence Engineering (MEIA)
Polytechnic of Porto – School of Engineering (ISEP)
Porto, Portugal
1222746@isep.ipp.pt

Mariana Carneiro
Masters in Artificial Intelligence Engineering (MEIA)
Polytechnic of Porto – School of Engineering (ISEP)
Porto, Portugal
1232152@isep.ipp.pt

***Abstract*—This article presents a continuous expansion of an ongoing project, focusing on anomaly detection and skin disease diagnosis using both image and text data. Additionally, it discusses the development of a chatbot to offer assistance to both patients and doctors. For image processing, models like VGG16, DenseNet, CNN, and ResNet were employed, while text processing and chatbot development utilized machine learning techniques such as LR, Naive Bayes, SVM, and deep learning models including Simple RNN, Conv1D, LSTM, BiLSTM, as well as specific architectures like BERT, LSTM, BiLSTM, SVC, and Transformers. The project's objective is to consistently enhance its capabilities, with the ultimate aim of delivering tailored support for various dermatological conditions.**

***Index Terms*—Skin Diseases Detection, Image Processing, Medical Chatbot, Digital dermatology, Early diagnosis, Natural Language Processing, Text classification**

I. Introduction

As the largest organ in the human body, the skin performs a fundamental role as a protective barrier. Its primary function is to protect the organism against harmful substances from the external environment while regulating the flow of essential nutrients [1]. However, daily exposure to several factors such as sun exposure, alcohol, smoking, viral agents, and exercise can negatively affect skin health and cause considerable damage, which can compromise the patient's health in some cases [2]. As a result, dermatological diseases are among the most prevalent health conditions affecting people across cultures and age groups, accounting for 30% to 70% of cases in high-risk populations [3]. Skin diseases not only affect daily activities and interpersonal relationships but can also significantly impact overall well-being and, in some cases, lead to death. Additionally, these conditions may contribute to mental health problems such as social isolation, depression and even suicide [4].

Considering the potential impact of such diseases, it is crucial to allocate more resources towards their treatment. In this context, early detection plays a significant role in reducing the impact of diseases and improving the overall well-being of patients and their survival.

This paper describes an approach to detect skin diseases in earlier stages using images and natural language to improve user involvement in this task.

II. State of Art

Advances in artificial intelligence (AI) are increasingly being used in the medical field, particularly in detecting skin diseases. The early diagnosis of dermatological anomalies, such as melanoma, lupus, psoriasis, urticaria and atopic dermatitis, has become a crucial priority for efficient treatment and prevention of severe complications. In this context, AI algorithms have played a fundamental role in allowing the precise recognition of cutaneous lesions through image analysis, predicting diseases through natural language text analysis, and developing conversational systems to clarify patient questions without the need for a doctor's intervention.

This state of the art aims to analyze recent advances in the field of computational dermatology, with a focus on the role of AI in the early detection of skin diseases and its potential to improve dermatological healthcare.

A. Deep Learning and Image Processing

Neural networks, inspired by biological nervous systems, learn patterns from data to automate problem-solving [5]. Similar to biological systems, they comprise artificial neurons processing inputs and applying transfer functions for outputs [6].

Techniques enabling deep neural network training, termed deep learning, significantly enhanced their efficiency [7], especially for modeling non-linear input-output relationships. Despite complexity, emphasis lies on connectivity between artificial neurons [8]. Deep learning encompasses various techniques, including 2D Convolutional Neural Networks (CNNs), vital in computer vision for image analysis [9]. CNNs efficiently process spatial and hierarchical patterns in images, employing convolutional layers to extract features and pooling layers to reduce dimensionality while retaining notable features. Fully connected layers then process these features for final decisions. Image pre-processing techniques, crucial in neural network data preparation, involve methods like image enhancement and histogram transformation [10]. These techniques ensure accurate extraction and representation of relevant information, enhancing model efficiency and precision [11].

Consider, for example, melanoma, an aggressive form of skin cancer. In recent years, there has been a significant increase in cases of malignant melanoma, which is responsible for tens of thousands of deaths each year in the United States alone [12]. Despite its lethality, melanoma is treatable, especially if diagnosed at an early stage. The survival rate can reach 96% if the abnormal proliferation of melanocytes is detected early, while in advanced stages the survival rate drops to 5% [13].

Significant progress has been made in the development of artificial intelligence algorithms for automatic or semi-automatic detection of neoplastic lesions based on analysis of optical images of individual dots [14].

A recent study presented a prototype whole-body imaging system that focused on evaluating the effectiveness of algorithms developed for skin lesion detection and segmentation. The combination of three algorithms, including deep learning methods and traditional approaches, results in an elevated level of sensitivity and precision in lesion detection. The geometric parameters of the segmented lesions were accurately calculated, especially for lesions with dimensions above 3 millimeters, which are the most suspicious for neoplastic [14].

Another study presented a prototype whole-body imaging system that focused on evaluating the effectiveness of algorithms developed for skin lesion detection and segmentation. The combination of three algorithms, including deep learning methods and traditional approaches, results in an elevated level of sensitivity and precision in lesion detection. The geometric parameters of the segmented lesions were accurately calculated, especially for lesions with dimensions above 3 millimeters, which are the most suspicious for neoplastic [14]. Moreover, the development of artificial intelligence has revolutionized the medical field, including the diagnosis of skin diseases. The use of advanced deep learning techniques enables accurate identification of dermatological anomalies through images, making a significant contribution to computational dermatology. The analyzed studies highlight the use of artificial intelligence models, such as AlexNet, VGG, GoogleNet and ResNet, to accurately identify skin diseases. The performance of these models has been improved using multi model fusion technology, which combines different deep learning approaches [15].

B. Natural Language Processing

Natural language processing (NLP) is an area of artificial intelligence that combines computational linguistics with statistical and machine learning models to enable computers to recognize, understand and generate natural language text [16]. In other words, NLP acts as a go-between, allowing natural language to be translated into a format that machines can understand.

NLP employes various techniques and approaches to address a wide range of linguistic problems. These have evolved from traditional methods based on statistical approaches to modern methods such as deep learning and neural network language models [17].

From a general perspective, systems that rely on parsing and semantic analysis have shown success in many tasks. However, given the vast amount of information available in machine-readable form, data-driven machine learning is assumed to be more effective in many cases. The two most used traditional language models are the n-gram model and the feedforward network, but both models have limitations in terms of context understanding. The n-gram model has a limited window of n-words, whereas the feedforward network has too many parameters and suffers from asymmetry in learning the position of a word in a sentence [18]. Nevertheless, some statistical models, such as Conditional Random Field (CRF), despite not being as prominent, still play a relevant role in specific tasks, including marking and classifying text [19].

On the other hand, Recurrent Neural Networks (RNN), including Long Short-Term Memory (LSTM), Sequence-to-sequence and Transformers, are the most used models for deep learning systems. In Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection, BERT, which stands for Bidirectional Encoder Representations from Transformers is a language model that is notable for its dramatic improvement over previous state of the art models. Furthermore, RNNs provide a useful approach for multiclass classification, where the vocabulary words serve as the classes. However, predicting the effectiveness of an RNN for a specific problem can be challenging without conducting experiments [18].

These technologies have played an important role in the healthcare industry, aiding in the generation of medical reports and clinical data analysis [19]. Additionally, they can be used to predict diseases based on non-structured data, which can be highly relevant in assisting doctors to make more accurate and precise decisions, minimizing their workload, and improving medical care.

A group of researchers has proposed a methodology that uses natural language processing techniques and deep learning to predict diseases based on unstructured data. The data is first transcribed into text using the Google API for speech recognition, then the text is pre-processed by removing stop words and tokenizing the words to be used by AI models. The proposed model comprises of an embedding layer that maps words to a vector space, an LSTM layer that captures long-term dependencies in the data, and a dense layer that processes the input and produces probabilities for each class. It achieved an accuracy of 98.94%, highlighting its significant advance in disease prediction tasks compared to the other models used [20].

In another study, researchers developed an innovative system for breast cancer prediction, combining machine and deep learning techniques. They utilized advanced feature generation methods, such as Bag-of-Words, Bag-of-CUIs, and structured data integration, to extract insights from clinical notes and patient information. Various machine learning classifiers, including random forest and logistic regression, were applied to these features, along with a knowledge-guided convolutional neural network (K-CNN) framework. Results showed significant improvements across configurations, particularly with the integration of word embedding, structured data, and CUIs, leading to an impressive F1 score of 0.5 for the K-CNN model [21].

Both studies employ distinct models and processing techniques. The first study utilizes natural language processing and deep learning models to transcribe unstructured data into text, achieving 98.94% accuracy. In contrast, the second study focuses on breast cancer prediction, employing advanced feature generation methods and a knowledge-guided convolutional neural network, resulting in a F1 score of 0.5. These findings underscore the importance of diverse approaches in healthcare predictive modeling.

Modern dialogue system architecture includes three main modules: NLP, dialogue manager, and natural language generation (NLG). The core of a dialogue system is analysis of user utterance inputted in NLP module [22]. Typically, in this module, the utterance is mapped to text vector representation (i.e., embeddings) [23]. Then vector representations are then used by the internal model to provide a response to the user. Chatbot could be considered intelligent if its responses are coherent and meaningful to the user. This behavior is highly dependent on the chatbot architecture and text vectorization methods.

Chatbots are divided into the following three categories based on the response generation architectures [24]: - rule-based chatbots, which analyze key characteristics of the input utterance and response to the user relying on a set of pre-defined hand-crafted templates; - retrieval-based (IR-based) chatbots, which select response from a large pre-collected dataset and choose the best potential response from the top-k ranked candidates; - generative-based chatbots, which produce a new text sequence as a response instead of selecting if from pre-defined set of candidates.

Latest works [25] show the high interest in generative-based chatbot architectures, thus rapid progress in this area is happening. However, it is worth noting that generative models require a huge amount of training data and computational resources while they are still likely to respond unpredictably which can be a genuine issue, specially in Healthcare. Therefore retrieval based chatbots continue to be more performant for Closed Domains, if comparing domain specific fine-tuned models such as BERT with Large Language Models [26].

Jennifer is an example of the successful implementation of AI-powered chatbots [27], highlighting their significant impact on addressing real-world challenges. Developed as a trusted source of information during the COVID-19 pandemic, Jennifer provided timely guidance and support for individuals seeking accurate updates and health-related advice. Her role exemplifies how AI can be leveraged for societal benefit, demonstrating the effectiveness of chatbots not only in healthcare but also in other domains where access to accurate information and timely assistance is crucial. While the specific types of machine learning models employed by Jennifer are not explicitly specified, her ability to interpret user queries and provide relevant responses underscores the versatility of AI-driven solutions, which can incorporate both rule-based approaches and machine learning algorithms, in facilitating information dissemination and supporting public health initiatives during crises.

Another widely recognized AI-powered tool is the "Ada - Your Health Guide" app [28], which has made significant strides in mental healthcare. By employing AI algorithms for retrieval, the app conducts adaptive interviews like medical assessments, providing users with probable diagnoses based on reported symptoms.The Ada model is based on a Bayesian network, built using extensive data from published clinical studies and input from expert clinicians, unlike most available SCs, which are not based on a machine learning approach [29]. Its effectiveness has been proven across various scenarios, offering a cost-effective and accessible means for early detection and referral to appropriate treatment.

Jennifer and the "Ada - Your Health Guide" app illustrate contrasting approaches to AI implementation in healthcare. While Jennifer utilizes a combination of rule-based and machine learning approaches without specifying the models employed, Ada relies on a Bayesian network model informed by extensive clinical data and expert input. This comparison underscores the diversity of AI-powered healthcare tools, with Jennifer showcasing versatility and Ada demonstrating the effectiveness of advanced machine learning techniques. Both highlight AI's potential to provide accessible and tailored healthcare solutions.

III. Development

A. Deep Learning and Image Processing

Exploring the technical aspects of deep learning implementation and image processing is fundamental to understanding the development of this project. In this section, the dataset used and the pre-processing process that was essential to prepare the data for analysis will be highlighted. These steps were crucial to ensure that deep learning models are trained with high-quality data that represent the dermatological conditions in question accurately.

1) Dataset and Pre-Processing

This study focuses on the analysis of a dataset consisting of images related to various dermatological conditions, such as urticaria, psoriasis, lupus, atopic dermatitis, and melanoma. In the training set, there are 212 images of urticaria, 1405 of psoriasis, 420 of lupus, 489 of atopic dermatitis, and 463 of melanoma. Meanwhile, the test set comprises 53 images of urticaria, 352 of psoriasis, 105 of lupus, 123 of atopic dermatitis, and 116 of melanoma. During the pre-processing stage, the first step is to verify and delete duplicate images in the dataset using hashing when importing the images. Secondly, all images are resized to a resolution of 250x250 pixels and subjected to linear segmentation and various filtering techniques. After that, the images were normalized. Additionally, to address class imbalance, oversampling and undersampling techniques are employed. The undersampling process involves two steps. First, reducing the majority class to the number of data points in the second majority class. Then, increasing the minority class to the same number of data points as the majority class. Afterwards, it was observed that the class situated in the middle differed little from the two majority and minority classes, so it was decided not to make any changes to it. After testing both approaches, undersampling was chosen due to the minority of urticaria images compared to psoriasis, the majority class.

2) Models and Modeling

"Regarding modeling, several convolutional neural network architectures were explored, including two-dimensional models, CNN2D, ResNet50, VGG16, and DenseNet. The initial goal was to obtain comprehensive results across all models, considering each dermatological condition as a class numbered from 0 to 5. Model tuning was conducted using both random search and grid search techniques to optimize hyperparameters. The results obtained were analyzed and compared using metrics such as accuracy, loss, confusion matrix, and graphs.

This study aims to contribute to the advancement of detection and classification of dermatological conditions through the application of image processing and deep learning techniques. The results obtained are crucial for the development of decision support systems in dermatology, aiming for more accurate and early identification of skin diseases."

B. Natural Language Processing

1) Text Classification

The dataset used in this part of the project contains 10334 rows with two columns: one for the disease label and another for the user's input describing their symptoms and how they are feeling. The description of the data may be visualized on table I.

The first step was to check if the dataset was balanced and the results showed that there were 2184 samples of Psoriasis, 2070 samples of Melanoma, 2216 samples of Urticaria, 1868 samples of Lupus and 2266 samples of Dermatitis. As the difference between the number of samples for each disease was not significant, it was considered that the dataset was balanced and there was no need to perform balancing techniques.

TABLE I
DISEASE CLASSIFICATION DATASET DESCRIPTION

Column Name	**Data Description**		
	Description	**Type**	**Values**
Disease	Identification of the disease	Categorical	Psoriasis Melanoma Urticaria Lupus Dermatitis
User_Input	Description of user's symptoms and how they are feeling	Text	-

Machine and deep learning models can struggle to comprehend natural language, including free text input. To overcome this limitation, a new column called **”User_Input_Preprocessed”** was created. This column contains a cleaned version of the user input achieved by removing stop words, single-letter words, digits, and special characters from the original text. Stop words are words that do not add significant information to a text, and eliminating them can speed up the text processing and improve model accuracy [30]. Additionally, the user input was standardized to lowercase to ensure consistency, thereby preventing the same word from being interpreted differently within the dataset.

After cleaning the data, a **WordCloud** was generated to visualize the most frequently used words for each disease. This helped to identify the key words that had the greatest impact on disease diagnosis.

To transform the sentences into a set of tokens the **WhiteSpaceTokenizer** from NLTK was used splitting the text by white spaces. Furthermore, stemming and lemmatization techniques were independently applied using **PorterStemmer** and **WordNetLemmatizer**, respectively, to reduce word variants to their root forms, known as lemmas. This approach reduces the number of unique words in the dataset and enhances the efficiency of subsequent modeling [31].

The implementation of the models was divided into two distinct parts, accommodating both machine learning and deep learning approaches. In the machine learning domain, models widely adopted for text classification due to their interpretability and robust performance were employed, including Logistic Regression, Multinomial Naïve Bayes, and Support Vector Machines.

To complement the preprocessing, the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques were employed for vectorization, transforming sentences into numerical vectors using **CountVectorizer** and **TfidfVectorizer**, respectively, with and without n-grams. The BoW technique represents a text as a collection of words, disregarding their order and structure. It involves creating a vocabulary containing all unique words in the dataset, and then each sentence is represented as a vector of word counts. However, BoW has limitations as it can generate large vectors and may overlook the relevance of some frequent words for the task. On the other hand, TF-IDF addresses the limitations mentioned above by evaluating the importance of a word in a document and across a collection of documents. Specifically, the importance of a word increases proportionally with the number of times it appears in the document but is offset by its frequency in the entire corpus. Therefore, TF-IDF considers both the local and global importance of words, resulting in more comprehensive evaluation of the words.

For each machine learning model, four different versions of the dataset were tested: one without any pre-processing, one with the cleaned version of the dataset mentioned above, and two based on the pre-processed version – one with stemming and one with lemmatization. Additionally, four types of word embeddings were used for each version of the dataset: Bag of Words, BoW including n-grams from 1 to 3 words, TF-IDF and TF-IDF including n-grams from 1 to 3 words.

To optimize the models, **GridSearch** and **Hyperopt** were employed for the Logistic Regression model, and **RandomSearch** was used for the Naïve Bayes and Support Vector Machine models. The **RandomSearch** was chosen as a second approach due to its speed when dealing with models that have a wide range of hyperparameters. These approaches helped to identify the best hyperparameters and achieve optimal results for all models.

Concurrently, advanced architectures were used for text classification in deep learning due to their capacity to capture complex patterns and dependencies within sequential data, such as Simple Recurrent Neural Network (Simple RNN), Convolutional Neural Network 1D (Conv1D), LSTM, and Bidirectional LSTM (BLSTM).

Initially, attempts were made to use deep learning models with different versions of the dataset and different types of word embeddings that were also used for machine learning. However, this approach proved to be time consuming due to the number of versions and the results were not significantly good. As a result, a different approach was used to deal with the text, using a **Tokenizer** to convert texts into sequences of numbers and then padding this sequence to convert all the inputs to the same length to complement the pre-processing. Then, for each deep learning model, four different versions of the dataset were tested: one without any pre-processing, one with the cleaned version of the dataset, and two others based on the pre-processed version - one with stemming and another with lemmatization.

Finally, attempts were made to improve the results of the models by changing the size of the embeddings as well as changing the architecture of the models.

2) Chatbot

When dealing with healthcare it is extremely important to be as accurate as possible, hence the reason the authors approach was a retrieval based chatbot. Retrieval based chatbots do not generate new utterances but they select an appropriate grammatically correct response from a large set of pre-collected Utterance-Response pairs. Given a dialogue context, both input utterance and responses pairs are encoded into some vector space representation, then the system counting semantic similarity score for each pair (i.e. dot product or cosine similarity) selects the best response from high-matched candidates. This approach based on information retrieval paradigm became popular in conversational agents [32].

The authors produced therefore a dataset of thousands of medical questions and answers and use pre-train models to find the best responses. Since the bot success depends on the quality of the dataset, the authors spent a significant amount of time researching and generating questions and answers to feed this information to the chatbot.

The chatbot accepts the diagnostic done from text classification model as intent and passes this value to a dialogue manager. Dialogue manager tries to find a more specific intent inside the main intent (it predicts the query purpose, such as treatments, or diagnostic), by using a BERT fine tuned with the dataset. The intent is predicted using transformer TFBertForSequenceClassification then a combined query is passed to the retrieval system including the customer query and the context (disease and intent).

Different models were experimented for text retrieval with impressive results when utilizing BERT Sentence Transformers from Hugging Face. The python framework Sentence-Transformers was used with a pre trained embedding model from Beijing Academy of Artificial Intelligence (BAAI) bge-small-en-v1.5, this embedding model was trained using Retro-MAE (Masked Auto Encoder) that utilizes the information of wikipedia and bookcorpus Hugging Face dataset. Together with a Sentence Transformer, a cross-encoder model was used (BGE Reranker) which uses a pre-trained cross-encoder model from Hugging Face Hub (bge-reranker-base) from BAAI. This cross encoder performs full-attention over the input pairs, and is more accurate than the embedding model, but requires significantly more time, therefore it only operates against the first top-k documents, re-ranking them.

Finally it was decided to combine the two models with a BM25 retriever, this implementing a Hybrid Retrieval (Keyword and Vector Search). BM25 considers both term frequency (TF) and document length normalization to determine a document's relevance to a given query. While vector search finds information conceptually similar, keyword search adds precision, improving the quality of the initial results [33].

Since all the turns are saved by dialogue manager, in case the patient asks a follow up question, it is possible to obtain the context and use it for adequate retrieval.

IV. Tests And Results

A. Deep Learning and Image Processing

1) CNN Model

The implemented CNN model consists of several convolutional and pooling layers, interspersed with dropout to prevent over fitting. Four convolutional layers were used, each followed by a max pooling layer, all with 3x3 filters and ReLU activation functions. The dropout rate was set to 0.5 after each convolutional layer for regularization.

After the convolutional layers, a global pooling layer was added to reduce the dimensionality of the extracted features. Then, two dense layers were included, one with 128 units and a ReLU activation function, followed by another dropout layer for regularization. The output layer has 5 units corresponding to the classes of dermatological conditions, with a soft max activation function. The results of this model are visible on table II.

TABLE II
RESULTS OF DISEASE PREDICTION (CNN)

Disease	Precision	Recall	F1-Score	Support
Atopic Dermatitis	0.37	0.40	0.38	123
Lupus	0.26	0.63	0.37	105
Melanoma	0.55	0.63	0.59	115
Psoriasis	0.70	0.33	0.45	350
Urticaria	0.19	0.23	0.21	53
True Positives (TP)	66			
True Negatives (TN)	49		Accuracy	0.42
False Positives (FP)	34		Macro AVG	0.40
False Negatives (FN)	9		Weighted AVG	0.43

2) ResNet Model

The ResNet model architecture begins with an input layer (250, 250, 3), resizing images to (224, 224). ResNet50 is used as the base model, excluding fully connected layers. After feature extraction, a global average pooling layer and a dense layer with units ranging from 64 to 512 are applied, followed by dropout. The output layer consists of a dense layer with 5 units for classes, using softmax for classification. Compilation uses the Adam optimizer with adjusted learning rate and categorical cross-entropy as the loss function.

The output layer consists of a dense layer with 5 units for classes, using softmax for classification. Compilation uses the Adam optimizer with adjusted learning rate and categorical cross-entropy as the loss function. The results of this model are visible on table III.

TABLE III
RESULTS OF DISEASE PREDICTION (RESNET)

Disease	Precision	Recall	F1-Score	Support
Atopic Dermatitis	0.44	0.28	0.34	123
Lupus	0.21	0.48	0.30	105
Melanoma	0.33	0.75	0.46	115
Psoriasis	0.65	0.33	0.44	350
Urticaria	0.00	0.00	0.00	53
True Positives (TP)	50			
True Negatives (TN)	34		Accuracy	0.38
False Positives (FP)	49		Macro AVG	0.31
False Negatives (FN)	3		Weighted AVG	0.37

3) VGG16 Model

The VGG16 model architecture is instantiated with an input layer of shape (250, 250, 3), and images are resized to (224, 224) to match the VGG16 input dimensions. The pre-trained VGG16 model is imported from the Keras API, excluding its top layer. The last layer of the VGG16 model, a GlobalAveragePooling2D layer, is removed to allow for custom output layers.

Following the base VGG16 model, a Flatten layer is applied to transform the output into a one-dimensional array. A dense layer with 256 units and ReLU activation is added, followed by a dropout layer with a rate of 0.5 for regularization. The final output layer consists of a dense layer with 5 units for multi-class classification using softmax activation. The resulting model is then returned as the VGG16 model architecture.

4) DenseNet Model

The DenseNet model architecture is defined with an input layer of shape (250, 250, 3). The DenseNet121 model is imported from the Keras API, excluding its top layer, and utilize it as the base model. The output of the DenseNet base model is then processed with a Flatten layer to transform it into a one-dimensional array. Subsequently, a dense layer with 256 units and ReLU activation is added, followed by a dropout layer with a rate of 0.5 for regularization. Finally, the output layer consists of a dense layer with a number of units equal to the specified number of classes (5 in this case) for multi-class classification using soft max activation.

The resulting model is then constructed with the input and output layers defined accordingly. This DenseNet model architecture provides a robust framework for classifying dermatological conditions based on input images. The results of this model are visible on table IV.

TABLE IV
RESULTS OF DISEASE PREDICTION (DENSENET)

Disease	Precision	Recall	F1-Score	Support
Atopic Dermatitis	0.42	0.20	0.35	123
Lupus	0.22	0.44	0.33	105
Melanoma	0.33	0.76	0.78	115
Psoriasis	0.67	0.35	0.43	350
Urticaria	0.20	0.31	0.42	53
True Positives (TP)	0			
True Negatives (TN)	0		Accuracy	0.44
False Positives (FP)	0		Macro AVG	0.28
False Negatives (FN)	0		Weighted AVG	0.28

These four models underwent constant modifications with the aim of maximizing accuracy. Various hyperparameters and their combinations were tested, explored through diverse techniques such as Random search, Grid search, and Hyperband Tuner, the latter of which was later discarded. Additionally, manual adjustments to the hyperparameters were necessary, as automatic adjustment techniques could lead to considerably long execution times, potentially reaching 10 hours or more of training.

Despite several attempts, the accuracy, recall and loss remained at similar levels. Consequently, exploring alternative solutions became necessary, such as modifying the data used in model training. This involved removing or altering the filters applied in pre-processing, adjusting image zoom, focusing on their centers, and varying batch sizes. Furthermore, both approaches to class balancing, namely oversampling and undersampling, as previously mentioned, were explored in the context of the models.

In summary, the rigorous experimentation process aimed to maximize accuracy, precision, and minimize loss to achieve optimal performance. All tests were meticulously documented. Due to their significant number, the authors opted to develop a script to review all models and select the best one. Ultimately, the CNN model with the architecture defined by the group emerged as the top performer. Additionally, after thorough research on Kaggle, the authors chose to explore a two-dimensional CNN model with the parameters previously defined to investigate potential issues with their architecture. However, upon executing the model and analyzing the results, the authors consistently found the outputs to fall short of expectations.

B. Natural Language Processing

1) Text Classification

For the disease classification were used machine and deep learning models that were trained for different kinds of pre-processing and word embedding. TableV shows the best results obtained for each model, including the best parameters and the pre-processing type.

TABLE V
CLASSIFICATION METRICS

Model	**Results**	**Parameters**	**Pre-processing**
Logistic Regression	Accuracy: 0.8911 Precision: 0.8942 F1-score: 0.8921 Loss: 0.3395	Solver: sag Penalty: l2 C: 8.768 Max_Iter: 10	Stemming + TF-IDF with N-gram
Naïve Bayes	Accuracy: 0.8623 Precision: 0.8628 F1-score: 0.8609 Loss:0.456	Fit_Prior: False Class_Prior: None Alpha: 0.430	Stemming + TF-IDF with N-gram
SVM	Accuracy: 0.8831 Precision: 0.8867 F1-score: 0.8843 Loss: 0.301	C: 0.1 Gamma: 0.1 Kernel: rbf	Stemming + TF-IDF with N-gram
Simple RNN	Accuracy: 0.854 Loss: 0.548	Embedding dim: 200	Stemming
Conv1D	Accuracy: 0.886 Loss: 0.594	Embedding dim: 100	Data cleaning
LSTM	Accuracy: 0.877 Loss: 0.462	Embedding dim: 100	Lemmatization
BILSTM	Accuracy: 0.864 Loss: 0.514	Embedding dim: 100	Stemming

Among the machine learning models, Logistic Regression stood out. Notably, stemming for text pre-processing surpassed lemmatization during testing, contributing to the model's improved performance in disease classification. Additionally, TF-IDF with N-gram embedding was found to be more effective than Bag of Words in capturing important patterns within the data. These specific pre-processing techniques played a crucial role in enhancing the performance of not only Logistic Regression but also the other machine learning models in disease classification.

All the models used for deep learning start with an embedding layer, which converts capture the relationships between words, and end with a dense layer that is suitable for making classifications. Moreover, all these models return a vector with the probabilities of an input being each on of the diseases. These architectures are tailored for different text classification tasks, offering varying complexities and capabilities to capture patterns in textual data effectively.

Between all the deep learning models tested, it was found that they were roughly equal in terms of accuracy. However, when the loss metric was taken into account, the best model was LSTM. Moreover, from these results it was seen that the embedding dimension of 100 performed well than the others , possibly due to not to long size of the padding sequences.

2) Chatbot

Since the models are pre-trained they have an official Benchmark MTEB (Massive Text Embedding Benchmark) [34], the models were selected for being state of the art, light and operate with Sentence Transformers. Therefore, the precision of the retrievals is extremely high. A valid answer from the Transformer to a query such as "What are the symptoms of melanoma?" would trigger a retrieval scored over 0.99. The authors tested over 90% of all queries related to the 5 diseases were answered correctly with a ranking score greater than 0.9. Only one question asked naturally out of 50 couldn't product a confidence > 0.6, that was a question about routine checks 'How often should I get my moles checked by a dermatologist for signs of melanoma?', this was due to the dataset not containing this information. The results of chatting with the bot depend on the dataset and questions asked, but since the domain is specific and narrow (5 diseases), it is obvious that the system will behave correctly without having to generate text, and since dealing with health, the system should be as exact as possible. Although the retrieval of the documents do seem extremely precise, the context of the query is obtain by the text classification model so one can consider the accuracy of the diagnostic paramount to the accuracy of the chatbot.

V. DISCUSSION OF RESULTS AND CONCLUSIONS

Despite efforts and tests conducted to predict one of the five dermatological diseases, the results always indicate a persistent trend of the model towards one of the existing classes. The difficulty in predicting one of the five dermatological diseases can be explained by their great clinical similarity, thus confusing the models, which need to identify subtle differences between them to predict accurately. This impasse highlights the importance of constantly improving the model and exploring more advanced approaches to accurately distinguish between the different dermatological diseases, even when they appear very similar. Adding to the difficulties were the computational resources, which posed a significant limitation for the development and training of the models. For instance, the limitation hindered the execution of the VGG16 model, adding complexity to the experimentation process. In spite of the challenges, the results unveiled the CNN as the most prominent model in terms of metrics.

When it comes to text classification, it's clear that the type of pre-processing and the type of embedding used in the text can have a huge impact on the performance of the models for classifying a particular disease. Despite the good results obtained in this task, it's clear that they could be improved by using a longer and more representative dataset.

Regarding the chatbot, it is clear that developing a fully fledge, production ready chatbot, is a arduous task, and would consume too much time and effort. Nevertheless the authors consider that this Information Retrieval chatbot is very accomplished, utilizes several modules that combined with its dialogue management capabilities is able to deliver great performance and usability. Even complex tasks such as multi turn management were at least partially accomplished. The authors also consider that there is much room for improvement, for example better classification of utterances, improvement of the dataset, further fine tuning of its models as well as sentiment analysis are features to be lacking. The dialogue manager could also interact with generative tools to, in certain cases, produce new responses, when the context is more generic.

References

[1] Shuang Cang, Yan Wang, Chengdu University of Information Technology, Institute of Electrical, Electronics Engineers. Chengdu Section, Institute of Electrical, and Electronics Engineers. *SKIMA : 2016 10th International Conference on Software, Knowledge, Information Management Applications : Chengdu University of Information Technology, China, December 15- 17, 2016*. IEEE, 2016.

[2] Nazia Hameed, Antesar Shabut, and M. A. Hossain. A computer-aided diagnosis system for classifying prominent skin lesions using machine learning. In *2018 10th Computer Science and Electronic Engineering (CEEC)*, pages 186–191, 2018.

[3] Angelo Picardi, Ilaria Lega, and Emanuele Tarolla. Suicide risk in skin disorders. *Clinics in Dermatology*, 31(1):47–56, 2013. Psychodermatology.

[4] Marie-Louise T. Johnson and Jean Roberts. *Skin Conditions and Related Need for Medical Care Among Persons 1-74 Years*. Number 212 in Vital Health Stat 11. Universidade da Califórnia, Nov 1978. PMID: 741665.

[5] Phil Picton. *What is a Neural Network?*, pages 1–12. Macmillan Education UK, London, 1994.

[6] P.D. Wasserman and T. Schwartz. Neural networks. ii. what are they and why is everybody so interested in them now? *IEEE Expert*, 3(1):10–15, 1988.

[7] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License. Last update: Thu Dec 26 15:26:33 2019.

[8] Jure Zupan. Introduction to artificial neural network (ann) methods: What they are and how to use them. *Acta Chimica Slovenica*, 41, 01 1994.

[9] Chunyan Yu, Rui Han, Meiping Song, Caiyu Liu, and Chein I. Chang. A simplified 2d-3d cnn architecture for hyperspectral image classification based on spatial-spectral fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:2485–2501, 2020.

[10] Tanzil Shahriar and Huyue Li. A study of image pre-processing for faster object recognition.

[11] Afonso Menegola, Michel Fornaciali, Ramon Pires, Flávia Vasques Bittencourt, Sandra Avila, and Eduardo Valle. Knowledge transfer for melanoma screening with deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 297–300, 2017.

[12] Howard W. Rogers, Martin A. Weinstock, Steven R. Feldman, and Brett M. Coldiron. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012. *JAMA Dermatology*, 151:1081–1086, 10 2015.

[13] Kenneth A. Freedberg, Alan C. Geller, Donald R. Miller, Robert A. Lew, and Howard K. Koh. Screening for malignant melanoma: A cost-effectiveness analysis, 1999.

[14] Michał H. Strzelecki, Maria Strąkowska, Michał Kozłowski, Tomasz Urbańczyk, Dorota Wielowieyska-Szybińska, and Marcin Kociołek. Skin lesion detection algorithms in whole body images. *Sensors*, 21, 10 2021.

[15] Ling Fang Li, Xu Wang, Wei Jian Hu, Neal N. Xiong, Yong Xing Du, and Bao Shan Li. Deep learning in skin disease image recognition: A review. *IEEE Access*, 8:208264–208280, 2020.

[16] IBM. What is nlp?

[17] Abhimanyu Chopra, Abhinav Prashar, and Chandresh Sain. Natural language processing. *International Journal Of Technology Enchancements And Emerging Engineering Research*, 1, 2013.

[18] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[19] Nan Luo, Xiaojing Zhong, Luxin Su, Zilin Cheng, Wenyi Ma, and Pingsheng Hao. Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal, 10 2023.

[20] Rahul Kumar, Tejaswi Rebaka, Jay Prakash, and Sushant Pradhan. Disease prediction from speech using natural language processing and deep learning method. *CIS 2020: Congress on Intelligent Systems*, 6 2021.

[21] Hanyin Wang, Yikuan Li, Seema A. Khan, and Yuan Luo. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artificial Intelligence in Medicine*, 110:101977, 11 2020.

[22] J.R. Bellegarda. Large-scale personal assistant technology deployment: the siri experience. In *INTERSPEECH*, pages 2029–2033, 2013.

[23] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2008.

[24] Ebtesam H. Almansor and Farookh Khadeer Hussain. *Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions*, pages 534–543. Springer International Publishing, Cham, 2020.

[25] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. In *arXiv*. 2020.

[26] Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about GPT-3 in-context learning for biomedical IE? think again. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[27] Ziang Xiao, Q. Vera Liao, Michelle X. Zhou, Tyrone Grandison, and Yunyao Li. Powering an ai chatbot with expert sourcing to support credible health information access, 2023.

[28] Stefanie Maria Jungmann, Timo Klan, Sebastian Kuhn, and Florian Jungmann. Accuracy of a chatbot (ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR Formative Research*, 3, 10 2019.

[29] Hamish Fraser, Daven Crossland, Ian Bacher, Megan Ranney, Tracy Madsen, and Ross Hilliard. Comparison of diagnostic and triage accuracy of ada health and webmd symptom checkers, chatgpt, and physicians for patients in an emergency department: Clinical data analysis study. *JMIR Mhealth Uhealth*, 11:e49995, Oct 2023.

[30] GeeksforGeeks. Removing stop words with nltk in python. https://www.geeksforgeeks.org/removing-stop-words-nltk-python/. Last Updated: 03 Jan, 2024.

[31] IBM. What are stemming and lemmatization? https://www.ibm.com/topics/stemming-lemmatization. Published: 10 December 2023.

[32] Abhay Prakash, Chris Brockett, and Puneet Agrawal. Emulating human conversations using convolutional neural network-based ir. *arXiv*, 1606.07056, 2016. Focus to learn more. Submission history: From: Abhay Prakash [view email]: [v1] Wed, 22 Jun 2016 19:55:24 UTC (574 KB).

[33] Microsoft Corporation. Hybrid search using vectors and full text in azure ai search. https://learn.microsoft.com/en-us/azure/search/hybrid-search-overview. Article. Published: 01/30/2024. Contributors: robertklee, HeidiSteen.

[34] Hugging Face. Bge-small-en model evaluation. https://huggingface.co/BAAI/bge-small-en#evaluation. Accessed on: 2024-04-11.