Prediction of Over-the-Counter Products and Exercise as Home Treatment Modalities for Knee Joint Pain Using Deep Learning

Abdurrhman Suliman
Department of Computer Science
Texas Tech University
Lubbock, Texas
absulima@ttu.edu

Chandrika Kancheti
Department of Computer Science
Texas Tech University
Lubbock, Texas
ckanchet@ttu.edu

Harry Onengiyeofori
Department of Computer Science
Texas Tech University
Lubbock, Texas
onharry@ttu.edu

Sruthi Manalapu
Department of Computer Science
Texas Tech University
Lubbock, Texas
srmandal@ttu.edu

VenkataMohanaRao,Nandigam
Department of Computer Science
Texas Tech University
Lubbock, Texas
vnandiga@ttu.edu

***Abstract*—Deep learning is utilized widely in the medical industry to detect and predict various outcomes. This paper focuses on utilizing a multilabel multiclass classification deep learning model to predict over the counter (otc) products, and exercises that a patient would use to treat knee pain. The project will be further extended to predict the user's post pain level after the selected treatment.**

Keywords—multilabel multiclass classification, deep learning, synthetic data generation, knee pain

I. Introduction

According to the National Library of Medicine [1] Knee joint pain affects approximately 25% of adults, and its prevalence has increased almost 65% over the past 20 years, accounting for nearly 4 million primary care visits annually. Although exercise and practice sport and reduce and delay knee pain, athletes who run or play sports that involve jumping or quick pivoting are also more likely to experience knee pain and problems [12]. Over time, it can be difficult to walk, run, use stairs, or engage in other activities requiring your legs. The good news is that several treatment options are available to reduce discomfort and help you get moving again. More than half of patients with knee joint pain utilize over the counter (OTC) pain medication. [2]. And Therapeutic exercise is often recommended as a means of treatment.[3].

Machine learning algorithms have demonstrated potential in swiftly and efficiently predicting both the likelihood of patients acquiring specific diseases or conditions, and the most appropriate treatments. As a branch of computer science, machine learning utilizes algorithms to detect patterns within extensive datasets and aids in forecasting various outcomes [4]. In numerous disciplines, machine learning techniques have become key instruments for prediction and decision-making. The availability of clinical data has significantly enhanced the role of machine learning in medical decision-making [6]. The development of a machine learning model could provide essential support, enabling real-time, effective decisions regarding treatments or exercises.

In this paper we aim to develop a hybrid approach that combines multi label deep learning models to accurately predict the most effective over-the-counter (OTC) medications and exercises for knee joint pain, using data collected from patient surveys. This model will leverage advanced neural networks to analyze patient characteristics and treatment outcomes, enabling personalized treatment recommendations.

Specifically, we are going to address the following research questions in this work:

Can Deep learning algorithms be utilized to predict the OTC Products and Exercise as Home Treatment Modalities for Knee Joint Pain, and if yes, what is the predicted post pain level the subjects felt after the treatment?

The next sections cover a review of the techniques that will be utilized in this project, the methodology implemented and the results we derived from the experiments.

II. Literature Review

Typically to carry out a prediction would require data. At the time of the project that was unavailable and so the project was updated to two parts, first to generate synthetic tabular data upon which the data can be modelled, and secondly to build

a model that can predict the treatments and post pain levels in one model.

We review methods for generating synthetic data, multilabel multiclass classification models for testing the efficiency of the synthetic data, and multi target deep learning model for predicting otc products, excersies and post pain levels.

Synthetic Tabular Data Generation: Data, and by extension synthetic data come in diverse types, from images to audio. For this project, the data is in tabular form. There are various methods for generating tabular data, but we review to main methods which can be divided into, the classical approaches and predictive approach [5]

The classical approach has two main methods
Baseline methods which utilize anonymization techniques and include "replacing data values, deleting sensitive attributes, and adding noise" [5]
ii. Statistical Models generate data by utilizing statistical and probabilistic models to simulate actual data and the relationships or correlations within the data [5]

The predictive approach also has two methods

i. Supervised Machine Learning Models which are trained on actual data and used to predict new records like the original records.[5]

ii. Deep Learning Approaches: In this case deep learning models are used to generate synthetic data. Two main methods include

    i. Autoencoders: An autoencoder comprises of an encoder, and decoder. The encoder compresses the data, creating and encoded representation of the real data, while the decoder decompresses the data so that it is a close variant of the original data.[5]

    ii. Generative Adversarial Networks comprise two neural networks, a generator and discriminator, that utilize an adversarial training process to generate synthetic data that is like real data.[5]

Since actual data is unavailable in this project, we utilize the Statistical Method for generating the dataset utilized. The method is explained in the methodology.

Multilabel Multiclass Classification: Prediction models fall into two categories, Regression models which predict numerical outcomes, and Classification models which predict probability of an outcome belonging to a given class.
Typical prediction models have one target feature, for classification models these include binary classification where one output has two classes, or multiclass classification in which the output has more than two classes. Classes are exclusive categories.

Extending the problem set the classes of an output variable could be further categorized, that is labeled into non-exclusive categories, for example a bird can be white or blue, but it cannot be pigeon and a dove at the same time. This leads to the model domain of multilabel classification, and they can be broken into two groups single label classification and multilabel multiclass classification models.

Below is a formal definition for multilabel multiclass classification.

Let X be a d-dimensional input space so that $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_3) \in X$ with,
output space of q with L labels
$L = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$, $q > 1$ so that $Y \subseteq L$ is called a label-set.
A multilabel dataset is defined by $D = \{(\boldsymbol{xi}, Y)) \mid 1 \leq i \leq m\}$
Given a quality criterion Q that rewards high prediction and low complexity. A Multilabel Multiclass Classification aims to compute the function $h: X \rightarrow 2^l$ such that h maximizes Q.

Multilabel Classification models can solve using the following methodologies

i. Problem Transformation Methods: These methods rely on transforming the dataset into a set of binary or multiclass problems, i.e., each dataset will have just one label. The main methods include
    a. Binary Relevance methods which transform the dataset in L binary classification problems, and [6]
    b. Label Powerset methods which creates a binary classifier for every label combination available in the dataset [7]
    c. Ensemble methods: Utilize a set of multi-class classifiers with methods like bagging and stacking to create a multi label ensemble classifier [6,7].

    Figure 1 is a picture of problem transformation methods.

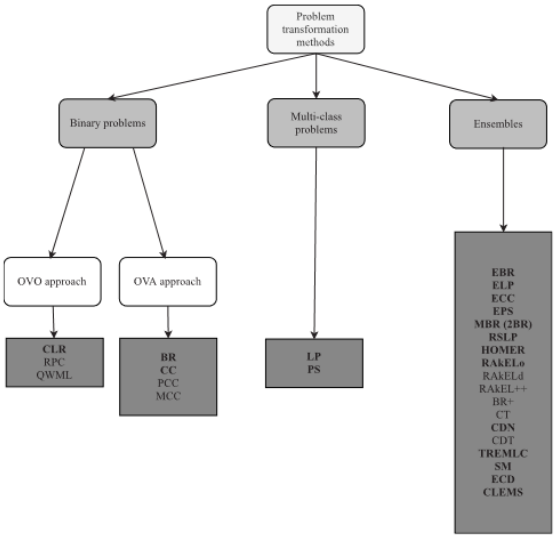

Fig 1. Problem Transformation Methods [6]

ii. Algorithm Adaptation Methods: They require adjustment to various machine learning and deep learning models so that instead of predicting one target variable, they can predict multiple. Figure II provides a description of such models and an explanation for an Artificial Neural Network follows

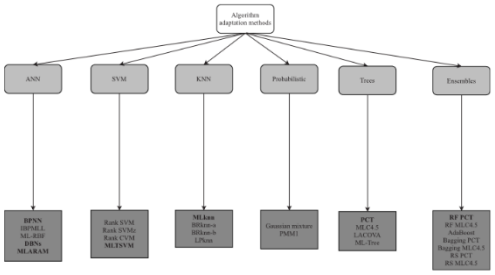

Fig 2. Algorithm Adaptation Methods [6]

Back Propagation Neural Networks (BPNN): This method utilizes a feed forward neural network with multiple output nodes in which each node represents a distinct label. It utilizes backpropagation to calculate the parameters of the network. Hyper parameters of the network are the learning rate, number of epochs and hidden units. Varying them has an impact on the performance of the model.

Our project utilizes a BPNN which we illustrate in the methodology.

Multitarget Learning: - write a comprehensive review

Multi-target prediction (MTP) is umbrella that covers several sub-areas of machine learning algorithms that share one major commonality, the simultaneous prediction of multiple targets of diverse type. These sub-areas include Multi-Output Regression, Multi-Label Classification and Multi-Output Classification. The goal is to predict multiple target variables simultaneously. These targets can be of different types, such as nominal, ordinal, or real-valued. This approach is beneficial in various domains where interdependencies exist among the targets and predicting them together can lead to better performance and more insightful models. In the lase years the interest in MTP has grown and the area has become more popular because many real-world problems need to predict multiple targets at the same time. Instead of predicting each number or category separately using methods that consider all outputs together can improve accuracy by taking advantage of the relationships between the targets.

Multi-target prediction (MTP) used in many areas include image tagging in computer vision, document tagging in text mining, and product recommendations in online retail. Other important MTP applications are found in climate science, where weather forecasting involves modeling relationships between atmospheric processes, and in medicine, where patients often have multiple interacting conditions. Additionally, in drug discovery, MTP methods can speed up finding chemical compounds that bind well with biological targets, which has been particularly crucial during the recent pandemic.

In this paper we are going to discover how multi-target prediction (MTP) is used to predict user preferences of over-the-counter products and exercises.

III. METHODOLOGY

Our problem requires building a multitarget deep learning model on synthetic data. The problem definition thus required that we split the methodology into two phases where Phase 1 would be to understand the data model, how to represent the dataset and build a baseline dataset and neural network model that we can test on. Phase 2 would then involve the actual construction of the dataset which would then be tested on a multilabel multiclassification models to determine the efficiency of the dataset and finally a multi-target deep learning model to predict selected otc products, accompanying exercises and post pain level from their application.

Phase I

Synthetic Data: While actual data was not made available, a survey template was provided from which we got the data model. The structure of the template has 91 features which are a mix of numerical and categorical data, and 6 output variables which can be broken into two groups – OTCSelect and ExcersiceSelect. Each OTCSelect question has 32 classes, while each ExcersiseSelect question has 53 classes.

From these we constructed the baseline data model as having 6 labels, 85 classes and 264 features. The output features represent all the classes, and each class can have a probability of 0 to 1, for each label assigned to each class. In counting the number of features, it was necessary to include a count of the classes of the categorical variables since they would be dummy coded. This gives us the result that data could be high dimensional for both X and Y, if the number of features is much greater than the number of observations. We decided to experiment with two types of synthetic datasets, high dimensional dataset, or wide dataset, and a long dataset in which the number of observations is much more than the number of features.

Utilizing the data model above, we generated the dataset using the "make_multilabel_classification" function from sklearn.datasets. This gives us the baseline dataset to test a baseline model.

Model: For the model we utilize a dense 4-layer BPNN. A picture of the layers is provided Figure 4. We utilize RELU as activation for the input and hidden layers, Sigmoid activation function as activation for the output label as explained above the sigmoid function ensures that each label has a probability between 0 to 1 that is assigned to each class. Binary cross entropy is utilized for the loss function.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 264)	69,960
dense_1 (Dense)	(None, 100)	26,500
dense_2 (Dense)	(None, 87)	8,787
dense_3 (Dense)	(None, 85)	7,480

Fig 4. Neural Network Architecture

To test each dataset, we utilize an 80/20 train-test split. We vary the number of epochs for the model on each dataset, for the long dataset we ran the model for 100 epochs utilizing a batch size of 100, and for the wide dataset we run for 200 epochs with a batch size of 20 using a 20% validation set.

Phase II

Synthetic Data: In this phase we begin the actual construction of the data set using the provided data model. Since we do not have any actual data, we can derive information from past research on knee pain, and construct data based on the conditional probabilities either provided or that we can derive from research. Fig 3 provides a flow chart for data generation.

First, we divided the dataset into two parts the demographic information and the knee pain survey data. The demographic data was sourced from freely available public data [9], [10]. The knee pain data, on the other hand, is not freely available. For that part we rely on provided information from past research to determine what fraction of our population should be positive for a specific condition. For example, to determine the cause of knee pain "OTCCause" we find from past research that some causes of knee pain are aging, and obesity [11], we also determine that a BMI above 25 is overweight [12], with that we can set all samples with BMI greater than 25 as cause obese, and all samples with age greater than 50 as aging.

Utilizing data from research in building out the features from the samples also helps us in having in-built relationships which can be modeled.

Model: This section is to be updated with multi-target learning model built.

IV. Results & Discussion

The results are broken down by the results from the phases

Phase 1:
The table below shows the results for both the datasets. Accuracy is used to measure the model's performance. From the results we infer from it that a neural network would work well in making predictions on the data model even if it is a high dimensional dataset. More critically we would need to ensure that the dataset built has relationships/correlations that can be modelled.

	(Accuracy)
Long Dataset	92.8%
Wide Dataset	93%

Phase II Result and discussion

1. **Exploratory analysis of the dataset with chi-square tests, and correlation test for the numerical**
2. **Model Results**
3. **Model Analysis**

Acknowledgment

We are thankful to Dr. Sheng for providing us the opportunity to work on his research project.

References

1- "Machine Learning Technique." ScienceDirect. https://www.sciencedirect.com/topics/computer-science/machine-learning-technique

2- S. J. M. Merchant, A. C. Li, A. P. Nguyen, M. N. Wirtzfeld, A. E. McLeod, J. S. Park, D. P. Dixon, J. N. Bathe, and A. B. Ball, "Clinicopathologic factors and adjuvant therapy in colonic

perforation," *Canadian Journal of Gastroenterology and Hepatology*, vol. 2016, Article ID 1360345, 6 pages, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5074793/.

3- C.-C. Wu, W.-C. Yeh, W.-D. Hsu, Md. M. Islam, P. A. Nguyen, T. N. Poly, Y.-C. Wang, H.-C. Yang, and Y.-C. (J.) Li, "Prediction of fatty liver disease using machine learning algorithms," *Computer Methods and Programs in Biomedicine*, vol. 170, pp. 23-29, 2019. doi: 10.1016/j.cmpb.2018.12.032.

4- R. Christensen, E. M. Bartels, A. Astrup, and H. Bliddal, "Effect of weight reduction in obese patients diagnosed with knee osteoarthritis: A systematic review and meta-analysis," *Ann. Rheum. Dis.*, vol. 66, no. 4, pp. 433-439, 2007. doi: 10.1136/ard.2006.065904.

5- M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, 2022. doi: 10.1016/j.neucom.2022.04.053.

6- J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, "Comprehensive comparative study of multi-label classification methods," *Expert Systems with Applications*, 2022. doi: 10.1016/j.eswa.2022.117215.

7- "Multi-label classification." Wikipedia. https://en.wikipedia.org/wiki/Multi-label_classification.

8- A. N. Tarekegn, M. Ullah, and F. A. Cheikh, "Deep Learning for Multi-Label Learning: A Comprehensive Survey," *IEEE Transactions on Neural Networks and Learning System*

9- "Describing Relationships Between Two Variables," PennState Eberly College of Science. https://online.stat.psu.edu/stat200/book/export/html/242

10- "Healthcare Insurance Dataset," Kaggle. https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance

11- "Knee pain - Symptoms and causes," Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/knee-pain/symptoms-causes/syc-20350849

12- "The Impact of Obesity on Bone and Joint Health," American Academy of Orthopaedic Surgeons. https://www.aaos.org/contentassets/1cd7f41417ec4dd4b5c4c48532183b96/1184-the-impact-of-obesity-on-bone-and-joint-health1.pdf

13- "The Impact of Obesity on Bone and Joint Health," American Academy of Orthopaedic Surgeons. https://www.aaos.org/contentassets/1cd7f41417ec4dd4b5c4c48532183b96/1184-the-impact-of-obesity-on-bone-and-joint-health1.pdf

14- "Management of Hypertriglyceridemia: Common Questions and Answers," *American Family Physician*, Nov. 1, 2018. [Online]. Available: https://www.aafp.org/pubs/afp/issues/2018/1101/p576.html.