**DETERMINATION OF GENUINENESS OF A MEDICAL PROFESSIONAL**

**(DOMAIN: MACHINE LEARNING, INFORMATION SECURITY)**

Aayush Shah Pranay Shah Dipam Vasani

Computer Engg Department, Computer Engg Department, Computer Engg Department,

D.J. Sanghvi College of Engg D.J. Sanghvi College of Engg D.J. Sanghvi College of Engg,

SAP – 60004140091 SAP – 60004140101 SAP - 60004140116

**ABSTRACT:**

In the age, where any data is easily available, it is of utmost importance that the data accessed by a person, isn’t used for malicious purposes, or more importantly the data isn’t in the wrong hands. This is even more important when the data is about an individual’s health and medical history. Often, we hear about cases wherein a medical professional misused a person’s past medical history. So, we thought of determining a medical professional’s genuineness, and then designing a system that may then decide to deny access to a malicious doctor. These results would be based upon the credulity of a medical professional. In this work, we discuss the design and implementation of the proposed project, using supervised machine learning algorithms to help determine whether a doctor is genuine or malicious. Factors like the location from which a medical professional accesses data, relevance of the data being accessed to the ailment/treatment, etc. would serve as inputs our deep learning model consisting of autoencoder and logistic regression algorithms. Our network would be trained on a dataset consisting of nine hundred thousand training examples. Finally, the genuineness of a medical professional would be obtained, and access would be granted or denied based on a predetermined threshold.

**PROBLEM DEFINITION:**

This main purpose of this project to find the credulity of a medical professional for a hospital management system. When a doctor accesses a patient’s data, we determine whether or not the doctor may misuse this information for any malicious purpose. Malicious doctors in hospitals often indulge in leaking a patient’s data to people outside the hospital. We can do this by looking at various cases such as the time and location of the data access. The access is likely to be genuine if the patient’s data is accessed within hospital hours and inside the hospital. Now, we can also refer to the type of data access and then map it to his profession. Ideally, a doctor will not access information that is not relevant to the area of his specialization. If he does so then there is a chance that he is malicious. We have used all these factors to determine whether the doctor is malicious or genuine. Data access can now be denied to a doctor who is found to be malicious.

**SCOPE:**

After extensive research that involved meeting professional medical practitioners and looking at medical research publications we have designed our rule base taking into consideration five major specializations. These include gynecologists, dentists, psychiatrists, dermatologists and ophthalmologists. Our system is designed to categorize malicious or genuine doctors that fall under these specializations.

Due to a patient’s medical records not being readily available online in the format we desire, we are designing our own data set for this project that is derived from similar datasets on kaggle. This will be done by referring a few sample data sets online and then forming our own data set. Thus, the neural network will be trained as per the examples from the data set develop by us. We have to go for this approach since data is not available freely in our desired form as no hospitals keep their data open source. If we get access to a data set in the form needed by us, we can use that to train our system.

A future scope of this project could be to increase the areas of specializations and perform additional research to accordingly update our rule base. The neural networks will have be trained again with examples from these additional specializations for it to find out the credulity of a doctor belonging to that profession.

**LITERATURE SURVEY:**

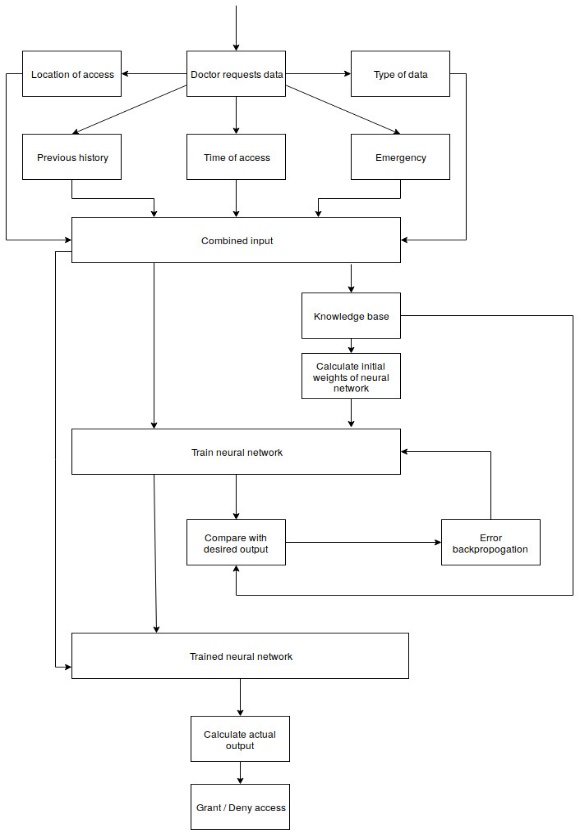
There are not many methods proposed by researchers that help restrict or in that case prevent any fraud against the misuse of a patient’s private data. While a lot of research has been done on protecting an IT system against unauthorized access from attackers, not much work has been done on preventing sensitive information from being jeopardized, either due to abuse or carelessness, by authorized users. And since it is extremely critical that we protect sensitive data against any kind of malicious intent, there is an increasing need for a more advanced, and highly secure and accurate system.

There was a proposed system [1] which used a quantified-risk adaptive approach in order to keep a check on the access of a patient’s privacy. The paper claims that even though doctors are authorized to access their patient’s data, in reality, there is also an inherent risk in each access. Their proposed solution therefore, allows information consumers to have the freedom to choose what they want to access. A user’s data-accessing activities are associated with quantified risk scores, which will be added up over time. Request to access a resource is granted if doing so will not make the user’s aggregated risk exceed his/her tolerance threshold set by the system; otherwise, the request is denied. They do this by calculating a relevance-relation function. They determine what activity a certain honest doctor will perform with respect to a disease, hence calculating the probability that a certain record m will be accessed to serve purpose p. They also take special cases into consideration. Once the model is trained they expect that a malicious doctor would over access a patient’s data hence increasing the risk considerably.

The problem with this method is that it is not as accurate as one expects it to be. Another glaring flaw is that they have only considered one parameter to determine if a doctor is malicious, i.e. relevance of the data accessed. The solution thus ends up ignoring myraid of other factors that can contribute towards determining spiteful activities, thus proving to be inefficient.

In another paper [2], clustering algorithms were used to analyse and hence group hospital data for better management. The paper mainly focuses on order history. It processes these order histories in order to find out the temporal global characteristics of clinical activities. Once it has done so, it keeps applying clustering techniques to the results until they converge. The final output is then expected to be the optimum output. For example the output may be a particular day when there is highest activity for a particular test or the time of the day when there is least activity. The major drawback of a clustering based approach is that, since it is unsupervised, the clusters formed may be the opposite of ones we expect. For example if there were a lot of malicious overprescribed orders in the history database, those will form a cluster and hence end up as a category.

**PROPOSED ARCHITECTURE:**

****

The process begins with a doctor requesting a certain data about a certain patient. The query is processed to find out the location of access, time of access and also the type of data. Whether the access is an emergency case or not is also mentioned by the doctor beforehand. All these, along with the previous history are combined to form an input.

If the network is not already trained the input is passed to the knowledge base so as to calculate the initial weights and also to the neural network as part of the training data. The neural network produces a certain output which is compared with the desired output (derived from the knowledge base) and the weights are adjusted accordingly. The loop continues till the network produces output as expected.

If the network is trained, the input is passed to the neural network and the actual output of the system is produced. This is again compared with a desired output, i.e. a threshold, hence producing a final output of the system, whether or not access will be granted. Finally the activity is logged into the previous history database for future reference.

**OUTCOME:**

We have successfully designed a system that determines the credulity of a medical professional and accordingly grants or denies access to him. We have used a supervised neural network for our problem statement that is trained by backpropagation technique to efficiently determine the result of our problem statement. We have created a well-defined knowledge base that helps our neurons in weight calculations and in determining the final output of our system. By doing this we aim to protect a patient’s confidential data in a hospital and prevent it from being misused for malicious purposes.

We have taken into account consideration for special cases and emergency situations in the hospital where data must be provided to a doctor as and when needed. Our rule base has taken into consideration five major specializations. These include gynecologists, dentists, psychiatrists, dermatologists and ophthalmologists. Our system is designed to categorize malicious or genuine doctors that fall under these specializations.

A future scope of this project could be to increase the areas of specializations and perform additional research to accordingly update our rule base. The neural networks will then be trained with examples from these additional specializations for it to find out the credulity of a doctor belonging to that profession.

**REFERENCES:**

[1] Quantified Risk-Adaptive Access Control for Patient Privacy Protection in Health Information Systems

[2] Clustering-based Analysis in Hospital Information Systems