Automated Approach to Selecting Neurological Medical Imaging Orders Using Natural Language Processing

Videet Mehta[1†], Rohan Dharia[2†], and Nilesh Desai[3,4]

[1]Math and Science Academy, Dulles High School, Sugar Land, Texas.
[2]Global Studies Academy, Travis High School, Richmond, Texas.
[3]Department of Radiology, Texas Children's Hospital and Baylor College of Medicine, Houston, Texas.
[†]These authors contributed equally to this work.

July 3rd 2023

Abstract

Medical imaging, like computed tomography (CT) and magnetic resonance imaging (MRI), holds profound value in disease diagnosis for millions worldwide. However, studies show that physician imaging orders may frequently be inappropriate (26% of cases) for the corresponding patient evaluation. Measures are necessary to mitigate patient risks in the subsequent re-imaging necessitated by physician error, including radiation exposure, additional sedation (pediatrics), and delayed treatment. To address these dangers, AIM-AI presents an unprecedented platform for automated medical imaging order selection using natural language processing and machine learning (ML). The algorithm was trained with anonymized imaging records and associated provider-input symptoms for 40,667 patients from Texas Children's Hospital, obtained after institutional review board approval. First, the data was preprocessed using tokenization and lemmatization to extract keywords. Second, an entity-embedding ML model converted the symptoms to high-dimensional numerical vectors suitable for model comprehension, which we used to balance the dataset through k-nearest-neighbor-based synthetic sampling. Third, a Support Vector Classifier (ML model) was trained and hyper-parameter tuned using the embedded symptoms to predict modality (CT/MRI), contrast (with/without), and anatomical region (head, neck, etc.) for an imaging order with 93.2% accuracy on 4,704 test cases. Finally, a web application was developed to package the model, which analyzes user-input symptoms and outputs the predicted order. The implementation of this application would save the lives of millions of patients facing potentially fatal risks associated with medical imaging by reducing costs, expediting treatment, and maximizing patient health. In this way, AIM-AI paves the path to a revolutionized medical field.

1 Introduction

Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are routinely used in the diagnosis of various diseases, playing a crucial role in disease identification for millions of people worldwide [1–3]. These imaging techniques allow healthcare professionals to visualize internal structures, enabling them to identify cancerous or diseased tissues, injuries, and neurological disorders. By using this information, medical professionals can carefully plan treatments to accurately address the underlying conditions.

1.1 Purpose

Although their general purpose is the same, MRI and CT scans are necessary for different diagnosis requirements. MRI is a non-radiation imaging process that utilizes the nuclear magnetic resonance phenomenon to generate images of soft tissue structures. By aligning hydrogen atoms in the body with the magnetic field and perturbing them with radio waves, the MRI scanner measures the resulting energy release and constructs a three-dimensional representation of the tissues being examined [2–7]. In contrast, a CT scan conducts numerous X-ray projections to create detailed cross-sectional images of the body with millimeter precision[8–10]. An intramedullary tumor, for example, would require an MRI scan due to its sensitivity in detecting soft tissue abnormalities. A broken bone, however, would require a CT scan due to its ability to visualize dense structures.

1.2 Imaging

The significant growth of the medical imaging industry has increased over two-fold from $3.6 billion to $7.6 billion over a period of 7 years[11]. This growth highlights the increasing prominence of medical imaging in healthcare. However, this rapid rise is also characterized by the need for improvements in imaging efficiency, from the initial physician encounter to the MRI/CT evaluation by the specialist. The imaging process involves numerous interactions between the patient, insurance company, physician ordering the image, and the clinic responsible for carrying out the order. As a result, the imaging process can lead to an extended patient timeline. A long process such as this in a rapidly increasing industry calls for the need for efficient and appropriate imaging. Artificial intelligence has proven to be necessary for CT scans for patient positioning, scan positioning, protocol selection, CT parameter selection, and image reconstruction [12].

1.3 Concerns

Imaging appropriateness has been of important concern in the effort to improve medical imaging efficiency and cost reduction. A past study showed that 26% of medical imaging orders were identified as inappropriate [13, 14]. This high proportion has four main consequences for patients. (i) Re-imaging leads to excessive radiation, which has a well-established link to cancer and other disorders [15–18]. (ii) MRI/CT scans require pediatric patients to be sedated, and repeating this can cause cardiopulmonary complications [19, 20]. (iii) The patient recovery timeline can be delayed due to the aforementioned long process of imaging having to be repeated, and for a patient with an

undiagnosed yet critical condition, timing can be the difference between life and death. (iv) The increased costs of reimaging can put a financial strain on not only the patient, potentially preventing them from further treatment, but also the hospital, which could better spend those funds to help more patients[21].

Physician-based clinical decision support systems (CDS) have been demonstrated to improve the efficiency of image ordering and reduce re-imaging rates significantly [22], and artificial intelligence has a potential impact in protocol/order selection in medical CT scans specifically [8]. However, limited research is available on the effectiveness of computer-based imaging order CDS algorithms. To address this gap in the literature, we conducted a study to develop an AI algorithm with NLP that can effectively determine the most appropriate imaging order for a patient based on their clinical signs and symptoms within a clinical setting. We hypothesized that implementing a computer-interface-based CDS would enhance clinical efficiency and reduce the likelihood of the four aforementioned devastating consequences of re-imaging a patient.

2 Methods

2.1 Data Collection

Data were obtained in CSV format from the Texas Children's Hospital Department of Pediatric Neuroradiology following IRB approval and HIPAA certification. All imaging orders were verified for accuracy and anonymized per HIPAA guidelines before their use in the study.

The dataset was composed of 40,667 entries of patient symptoms paired with the corresponding imaging order as seen in Figure 1. Exploratory data analysis [23] was then conducted to identify key features for implementation in the algorithm. The initial dataset consisted of approximately 80 neurological imaging orders but was eventually reduced to the top 8 most frequent imaging orders. The distribution of the frequency of imaging orders in Figure 2 reports CT Head without Contrast as a frequent option or ordering and MRI Pituitary with less than 1000.

The data in Figure 1 included information on various unidentifiable patient attributes namely age, weight, and height. which were assessed for their contribution to the model's output. It was determined that these columns would introduce unnecessary features with low predictive power due to their randomness and were therefore excluded. The average word count of the clinical signs/symptoms showed an average of 4.6 words indicating the limited information provided with each entry (Fig. 3). Three key issues were identified in the data that impacted our approach prior to algorithm implementation. First, the degree of irregularity present required the implementation of a pre-processing step prior to ML. Many rows consisted of abbreviations, irregular capitalization, incorrect spelling and grammar, and empty entries. Second, a severe data imbalance, with nearly 11,000 cases of CT Head without Contrast and just 1,000 of MRI Pituitary indicated the need for class balancing, either with under-sampling or oversampling [24]. Lastly, the most frequent word count of the symptom text rows was found to be just three, as seen in the distribution graph, indicating that each word present contributes significant meaning and that any pre-processing should be somewhat conservative (Fig. 3).

2.2 Preprocessing

A three-step processing algorithm addressed these data irregularities. The initial step was to normalize and condense the symptom text. Our initial approach was to utilize OpenAI's Davinci GPT-3 API [25] to ask it to extract important keywords and include similar words to the existing prompt for additional data. This approach proved to be costly and computationally expensive as it required API and tokens for purchase. The alternative was the Natural Language Toolkit (NLTK) which tokenizes each symptom row and filters through the StopWord corpus to identify stop-words [26]. As opposed to extracting important keywords, the algorithm uses the corpus to identify and extract unnecessary filler words that serve no significant meaning such as "with", "to", and "also". It proved to be the most effective approach in normalizing the data and accounted for special characters and irregular syntax as seen in Figure 5. However, using the NLTK library sacrificed the ability to correct spellings, a feature that the GPT model would automatically account for.

Following keyword filtering, word lemmatization was performed to simplify each word to its most fundamental grammatical root and maintain consistency across conjugations as seen in Fig. 5 (e.g. "flying" transformed to "fly") [27, 28]. Lemmatization was preferred over stemming due to its higher accuracy but at the expense of computational speed [29, 30]. Without lemmatization, however, the model is introduced to more variability in words and would impact the accuracy of the model.

To facilitate ML model training, it was necessary to transform the symptom text into numerical representations [31]. The reasoning of use of an embedding model for text-to-numerical representations was two-fold over the established one-hot encoding approach - the computation time is significantly increased and the dimensionality of the encoding approach of a 40,000-row dataset would magnify the complexity of the algorithm carrying out downstream tasks [32, 33]. The focus went on utilizing the optimal model for entity embedding. With limitations of computational power and a limited solution timeline, the choice was dependent on the previous literature on the trade-off of accuracy to time complexity. We leveraged OpenAI's 1.2 billion-parameter pre-trained and custom-tuned Ada embedding model [34], which uses entity embedding techniques to convert each symptom row into dense 1536-dimensional numerical vectors to convey both the definition and context of the input (Fig. 5). The model architecture first embeds each word into a vector and then uses multi-head self-attention transformer layers to understand and generate relationships between the words. Word2Vec, a prospective alternative embedding algorithm, was not preferred in this use case due to possible words un-encountered in the pre-trained base embedding and presence of pairs of structurally similar words but different definitions [35].

Referenced before, Figure 2 magnifies the severe class imbalance present in the dataset. Training an ML model with imbalanced data with a majority-minority ratio of approximately 15:1 is sub-optimal to maximize accuracy as much as possible. To combat this, a synthetic sampling algorithm assisted in balancing the dataset following entity embedding. We utilized a synthetic minority oversampling technique (SMOTE) k-nearest-neighbor algorithm [36], which is designed to construct new points close to existing ones in high-dimensional feature space, thereby increasing the number of samples of the minority class. A sample point is paired with a neighbor and the difference between the points is multiplied by a random number which then creates a new point in between

the two existing points. SMOTE has also been proven to work well in high-dimensional datasets [37]. Each of the minority classes was oversampled to match the count of the majority class (CT Head without Contrast). SMOTE increased the sample size from 40,667 to approximately 90,000 rows and was used to train the final ML model (Fig. 7). Basic minority oversampling and majority undersampling techniques were also tested, however, the SMOTE algorithm was found to be more effective at increasing the variety of data present in a limited dataset [38]. By generating synthetic data points, the algorithm was able to maintain the underlying structure of the data while eliminating the problem of class imbalance.

2.3 Machine Learning

Selecting a model required a rigorous experimental design, training, and testing of three distinct supervised alternatives on the same dataset with default model parameters. This approach allowed for a reliable assessment of the most effective model for accurate classification. Specifically, the models tested were XGBoost, which utilizes gradient boosting and decision trees to iteratively learn from residuals and minimize loss; Random Forest, an ensemble learning method that constructs multiple decision trees to aggregate their outputs and improve prediction accuracy; and Support Vector Machine (SVM), a robust classification algorithm that identifies a hyperplane to optimally separate data points of different classes in a high-dimensional feature space through a maximum margin function. Evaluation of all three models proved the Support Vector Classifier to be the best for the dataset as seen in Figure 8. The reasoning of the model was two-fold. (i) It was critical to prevent the model's overfitting, a problem commonly associated with high-dimensional data SVMs [39, 40]. The SVM is comparatively better than regression algorithms and other classifiers in preventing overfitting in high-dimensional space leading to overall improved performance in high-dimensional datasets [41, 42]. (ii) SVM classification models being resistant to noise [40]. This is important in medical NLP where abbreviations and other anomalies are included in the history [43].

The straight line for the boundaries of each classification region in a linear SVM classifier is defined as:

$$\beta_0 * x_1 + \beta_2 * x_2 = -\beta_0 \quad (1)$$

where $\overrightarrow{w}$ is $[\beta_0, \beta_1]$ which is perpendicular to the hyperplane and $\overrightarrow{x}$ is $[x_1, x_2]$ is the point on the straight line.

Equation 1 can be generalized to:

$$\beta_0 + \beta_0 * x_1 + \beta_1 * x_2 + ... + \beta_p * x_{p+1} = 0 \quad (2)$$

or

$$\overrightarrow{w} \bullet \overrightarrow{x} + b = 0 \quad (3)$$

X_i is defined as the vector to display all features of index i. With $\overrightarrow{X_+}$ describing the positive class and $\overrightarrow{X_-}$ describing the negative class. The support vectors can be defined as

$$\overrightarrow{w} \bullet \overrightarrow{X_+} + b > 1$$

and

$$\overrightarrow{w} \bullet \overrightarrow{X_-} + b < -1 \quad (4)$$

Provided that Y_i is a vector containing values -1 and +1. Equation 4 can be generalized into

$$Y_i(\overrightarrow{w} \bullet \overrightarrow{X_i} + b) \geq 1 \quad (5)$$

The width of the margin is then defined as

$$\frac{(\overrightarrow{X_+} - \overrightarrow{X_-}) \bullet \overrightarrow{w}}{\|w\|} \quad (6)$$

or

$$\frac{2}{\|w\|} \quad (7)$$

And to maximize Equation 7, the model must minimize

$$\frac{1}{2} \|w\|^2 \quad (8)$$

Lagrangian Transform then optimizes Equation 8 with constraints of Equation 5.

$$L = \frac{1}{2} \|w\|^2 - \sum \alpha_i [Y_i(\overrightarrow{w} \bullet \overrightarrow{X_i} + b) - 1] \quad (9)$$

Since

$$\frac{\delta L}{\delta w} = 0, \text{ therefore } \overrightarrow{w} = \sum \alpha_i Y_i X_i \quad (10)$$

and

$$\frac{\delta L}{\delta b} = 0, \text{ therefore } \sum X_i Y_i = 0 \quad (11)$$

Which leads to

$$L = -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j Y_i Y_j \overrightarrow{X_i} \bullet \overrightarrow{X_j} + \sum \alpha_i \quad (12)$$

From here, we see that the SVC depends on $\overrightarrow{X_i} \bullet \overrightarrow{X_j}$.

Each model was trained on the pre-processed data and was validated on a 4704(5%) randomly sampled test case set with accuracy metrics. Accuracy, in terms of True Positive(TP), True Negative(TN), False Positive(FP), and False Negative(FN).

$$Accuracy = \frac{TP+FP}{TP+FP+FN+TN} \quad (13)$$

After selecting a model, the support vector machine further optimized the model's performance through GridSearchCV which further optimized the model's performance by conducting tuning of six model parameters using GridSearchCV. The GridSearchCV starts with random hyperparameters and slowly optimized the model through a trial-and-error method. The perfect model is met when there is a balance between underfitting and overfitting, therefore increasing the metrics of the model [44–46]. The parameters that were optimized were Kernel, C, Gamma, and Train-Test split. A low-degree polynomial kernel is optimal in NLP situations [47, 48].

With x and y reprinting the input vectors, $\varphi(x)$ and $\varphi(y)$ representing the mapping functions, the polynomial kernel works as follows:

$$K(x, y) = (x^T y + c)^d = (\varphi(x), \varphi(y)) \quad (14)$$

3 Results

The models were successfully trained and tested on the preprocessed dataset. Notably, the SVM exhibited the highest accuracy of 91.5% on a randomly sampled test case set of 4704 from the original data, outperforming the other models (see Fig. 8 for model accuracies). The GridSearchCV then increased the SVM accuracy by 1.7% to 93.2

Precision, recall, and F1 scores painted a more detailed picture of the intricacies of the model (Fig. 9), and each can be defined in terms of the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) cases for the given classification in the test set.

Precision represents the proportion of retrieved positive cases that were relevant and is calculated as:

$$Precision = \frac{TP}{TP+FP} (15)$$

Recall represents the proportion of relevant positive cases that were retrieved by the model and is calculated as:

$$Recall = \frac{TP}{TP+FN} (16)$$

The F1 score is the harmonic mean of precision and recall.

$$F1 = \frac{Precision*Recall}{Precision+Recall} (17)$$

MRI Pituitary WO/W Contrast had the highest scores of all three metrics at 0.97, 0.99, and 0.98 respectively, suggesting confidence with this class. MRI Brain WO/W Contrast had the lowest scores of all three metrics at 0.84, 0.72, and 0.77 respectively, suggesting that this class may be highly similar to another. However, the average of all individual class metrics came out to be 0.92, 0.92, and 0.92 respectively, confirming the model's overall accuracy.

The heatmap in Fig. 10 assists in visualizing class-specific performance as well. Similarly, the receiver operating characteristic (ROC) curves seen in Fig. 11 demonstrate a strong tradeoff between true and false positives. An area under the curve (AUC) close to 1 for each class indicates an accurate model. The model achieved near-perfect AUC values for three classes: CT Soft Tissue Neck With Contrast, CT Maxillofacial Without Contrast, and MRI Pituitary WO/W Contrast. These findings are consistent with the high metrics observed for the classes above in Fig. 9.

To properly visualize the internal workings of the algorithm, the 1536-dimensional data was compressed into a two-dimensional array through a t-distributed stochastic embedding (TSNE) network. tSNE is a dimensionality reduction technique that embeds the 1536 features present into two or three dimensions for visualization. Each individual point represents a case, all of which can be observed to be clustered based on class.

In order to carry out the tSNE algorithm, the high-dimensional Euclidean distances between any two datapoints x_i and x_j is first found in the form of conditional probabilities, which represent the similarity between the points. The conditional probability $p_{j|i}$ that x_j is near x_i is calculated using a Gaussian distribution centered at x_i with a standard deviation of σ_i and is defined as:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)} (18)$$

The high-dimensional joint probability distribution can then be calculated as:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \quad (19)$$

Next, a second joint probability distribution is constructed in low-dimensional space. The conditional probability q_{ij} representing the similarity of any two points y_i and y_j is calculated with a t-distribution:

$$q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1+\|y_k - y_l\|^2)^{-1}} \quad (20)$$

Finally, the deviation between the low- and high-dimensional data point distributions must be minimized, which can be done using the Kullback-Leiber (KL) divergence. The KL divergence for distributions P and Q in space χ is defined as:

$$D_{KL}(P \,||\, Q) = \sum_{x \in \chi} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (21)$$

Gradient descent is employed to iteratively modify the low-dimensional data to best match the high-dimensional data. To do this, the cost function C must be minimized. C represents the KL divergence of the joint probability distributions P, from the high-dimensional data, and Q, from the low-dimensional data and is defined as:

$$C = KL(P \,||\, Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \quad (22)$$

The results are shown in Fig. 12.

Hyperplanes were then drawn to view the classification boundaries of the SVM (Fig. 13). Due to its large compression ratio ($1536/2 = 768$), many features were most likely lost, so the classification in a two-dimensional space is not entirely representative of the 1536-dimensional hyperspace but does serve as a rough model. The poor classification of points seen in the MRI Brain WO/W Contrast class in Fig. 13 is consistent with the relatively low class-specific metrics observed in the classification report in Fig. 9.

The accuracy of the SVM model does appear to be significantly higher than the physician accuracy of 74.3%, but a one-way two proportion z-test confirmed that there was a statistically significant difference in the physician imaging order accuracy of 74.3% and SVM model imaging order accuracy of 93.2% ($p < 0.0001$). In the context of the 100 million MRI/CT scans done annually, an increase of 18.9% in imaging accuracy can prevent 18.9 million patients from the effects of excessive imaging.

In order to maximize the efficacy of the ML model and provide a more realistic clinical tool, a probability-based prediction system was implemented that provides the probability of each class fitting the input, rather than outputting a single classification. The algorithm, by default, offers the top three highest-probability class results with a 100% chance that the class labeled as correct will be present in the top three (Fig. 14).

4 Discussion

This algorithm proves the effectiveness of an automated approach to mitigating existing risks of incorrect imaging order selection by physicians. However, a raw algorithm is impractical for a

practicing physician, who requires a straightforward user interface. After the study, the model was packaged into a web application that receives free-text symptom input and uses it to output the predicted imaging order with a confidence score. This application acts as a clinical decision support tool, allowing physicians an additional verification step to produce a combined imaging order selection accuracy of over 93%.

A notable use for this tool is in emergency settings. Rather than requiring a physician to evaluate a patient both before and after receiving imaging, the application would reduce strain on the physician by eliminating any need for them to see the patient until after imaging and diagnostics have been completed, which would maximize hospital efficiency.[49] Patients could realistically be rapidly evaluated, receive an imaging order from the tool, undergo imaging, and have the results read in a short time frame, which would improve their treatment outlook.

The widespread implementation of this algorithm holds potential for several positive outcomes as a result of decreased re-imaging due to more accurate imaging order selection.[22] First, excessive radiation brought by several imaging methods, namely CT scans, would be mitigated. Presently, radiation, which is significantly carcinogenic, can be presented in unnecessarily high dosage when imaging must be carried out multiple times.[15, 50] With improved imaging order selection, fewer cases of re-imaging would reduce radiation exposure to a minimal level.

The algorithm would, furthermore, eliminate repeated sedation in patients, reducing the potential for complications. Pediatric patients, as well as adults with certain conditions such as claustrophobia, require sedation when being imaged, but such practices pose risks of complications for those patients.[51] By minimizing the number of instances of sedation, the algorithm will maximize patient health outcomes.

Reducing re-imaging would also expedite patient treatment timelines. The process of imaging, including carrying out the imaging, reading results, and conveying the outcome to the patient, is time-consuming, and for patients with unknown critical conditions, repeating these steps due to physician imaging order error can be life-threatening. Without re-imaging, the duration of diagnostic testing is lessened for improved treatment options. [52].

A final advantage is the reduction of costs for both the hospital and patient that is brought by more precise imaging ordering. [53] Hospital systems currently spending millions to carry out imaging orders would see significant drops in spending with the implementation of this algorithm.[11] By instead applying these funds to causes such as research, medical facilities could see a growth in innovation and potentially revolutionary treatment breakthroughs.

Despite the numerous benefits of this tool, there exist several avenues for further development that would optimize its effectiveness. Enhancement of this algorithm's accuracy metrics will most importantly require the incorporation of more training data. The present data includes just one feature (clinical symptoms) and originates from a single hospital institution. Incorporating vital signs, laboratory orders, family history, and clinical guidelines from multiple institutions in the model could certainly see an additional 4-5% accuracy increase.

The present dataset also is limited to neurological imaging orders; however, due to the generalizability of the algorithm, this can be easily diversified.[54] Because this model does not need to be adapted to specific features of the dataset, any additional, similarly-structured data (free-text

input + class-based output) could be used in training to also produce high accuracy rates. In this way, with datasets for other body systems, such as gastrointestinal imaging orders, the scope of the model could quickly be grown to fit any medical specialty, and the application's online nature will allow for real-time remote updates to the platform.

5 Conclusion

This algorithm has significant potential to be truly revolutionary for the millions of patients facing incorrect imaging orders in the United States. Beyond this, scaling the application to the rest of the world would impact millions more by reducing costs through fewer scans, expediting treatment through more efficient diagnosis, and maximizing patient health through reduced radiation and sedation. This tool will certainly be critical for a more efficient healthcare system and groundbreaking to the medical field as a whole.

6 Figures

PHYSICIAN IMAGING ORDER	AGE AT ORDER IN YEARS	WEIGHT IN LBS	HEIGHT IN FT INCHES	CLINICAL SIGNS/ SYMPTOMS
CT HEAD WITHOUT CONTRAST	17.58	178.13188	5' 7.717"	abnormal gait, word finding difficulty, tremor
CT SOFT TISSUE NECK WITH CONTRAST	16.38	90.829	5' 1.1417”	trismus, uvular deviation, muffled voice - concern for deep neck infection vs PTA
MRI FACE WO/W CONTRAST	17.91	216.49188	5' 8.661"	right ear vascular mass; possible AVM

Figure 1: Sample Data

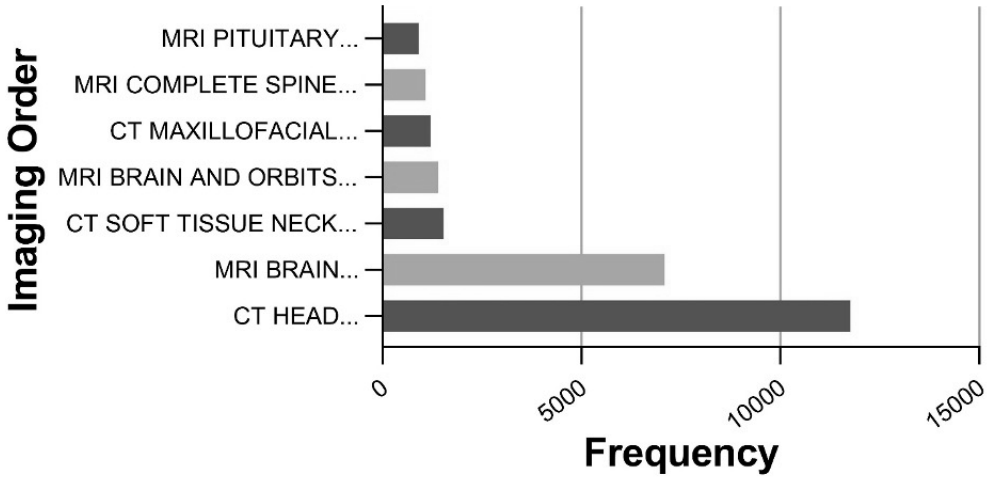

Figure 2: Class Distribution

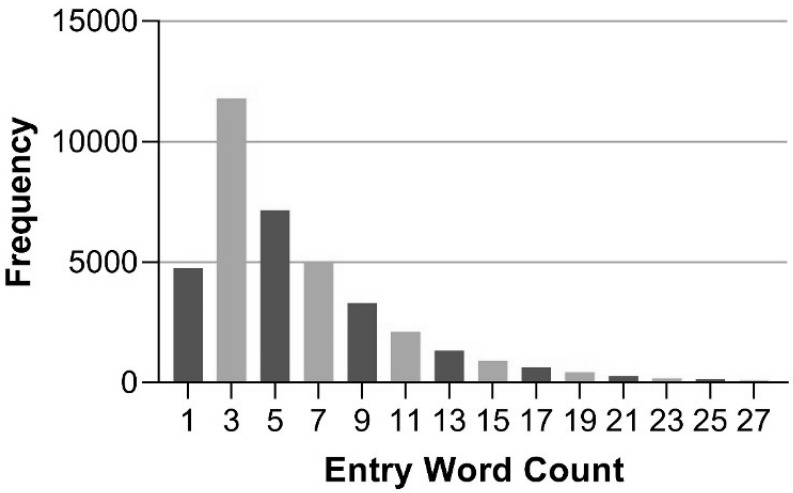

Figure 3: Word Count Distribution

Abnormal	Gait	Word	Finding	Difficulty
Rapidly	Enlarging	Right	Neck	Swelling
Lethargy	Prolonged	Sleeping	Evaluate	Encephalopathy

Abnormal	Gait	Word	Find	Difficulty
Rapid	Enlarge	Right	Neck	Swell
Lethargy	Prolong	Sleep	Evaluate	Encephalopathy

Figure 4: Lemmatization

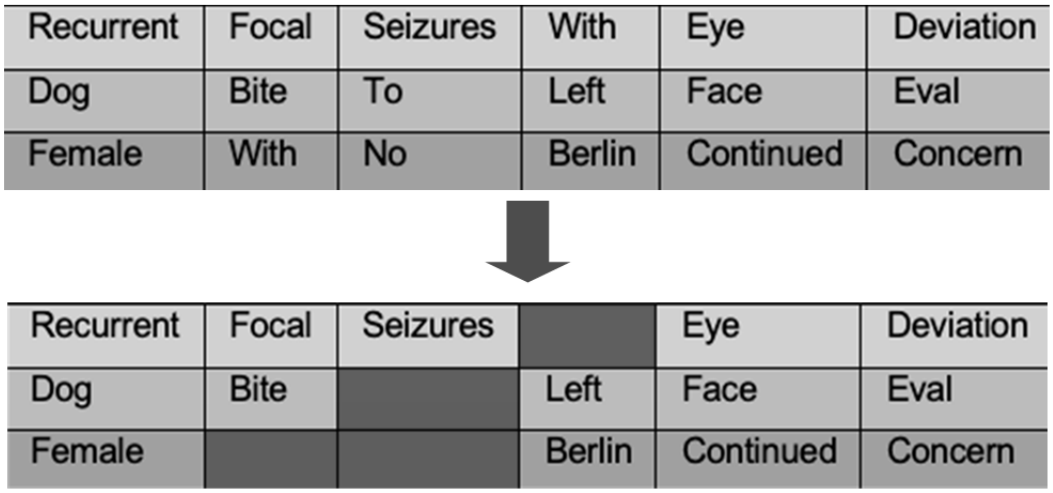

Figure 5: Keyword Filtering

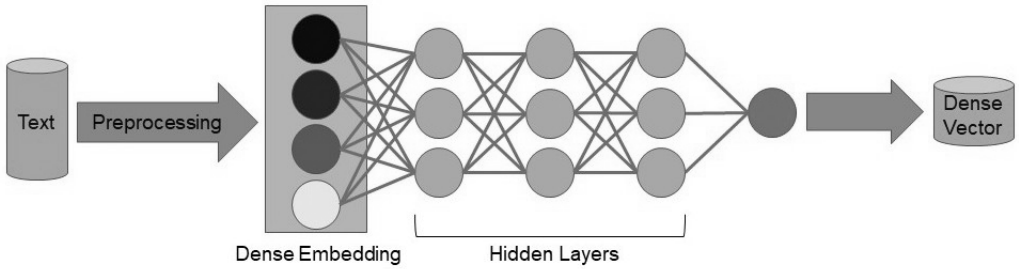

Figure 6: Entity Embedding

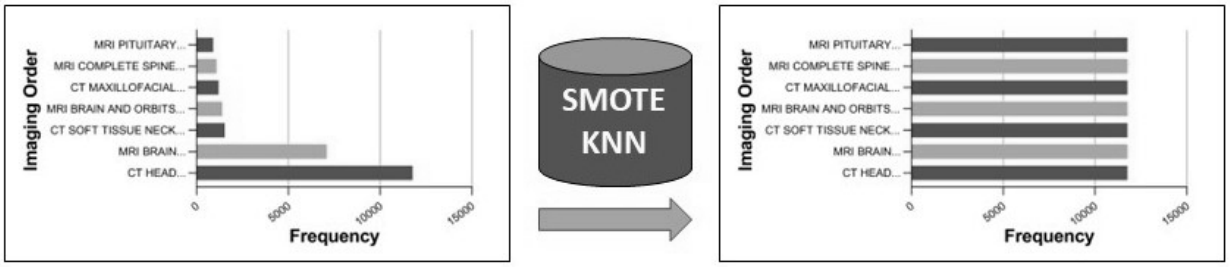

Figure 7: Synthetic Sampling

Model	Description	Accuracy
XGBoost	Gradient boosting decision tree optimized for efficiency that uses regularization, parallel processing, and data structure optimization	84.6%
Random Forest Classifier	Ensemble machine learning algorithm utilizing multiple decision trees, each employing a unique criterion for recursive partitioning of the feature space and compiling a prediction	82.3%
Support Vector Classifier	Discriminative classification algorithm that finds an optimal hyperplane that maximally separates the label classes in high-dimensional feature space	93.2%

Figure 8: Model Comparisons

Imaging Order	Count	Precision	Recall	F1
CT HEAD WITHOUT CONTRAST	585	0.85	0.83	0.84
CT MAXILLOFACIAL WITHOUT CONTRAST	588	0.95	0.99	0.97
CT SOFT TISSUE NECK WITH CONTRAST	577	0.97	0.99	0.98
MR ANGIO BRAIN WITHOUT CONTRAST	585	0.94	0.98	0.96
MRI BRAIN AND ORBITS WO/W CONTRAST	603	0.94	0.94	0.94
MRI BRAIN WO/W CONTRAST	557	0.84	0.72	0.77
MRI COMPLETE SPINE WO/W CONTRAST	617	0.91	0.95	0.93
MRI PITUITARY WO/W CONTRAST	590	0.97	0.99	0.98
Sum/Average	4704	0.92	0.92	0.92

Figure 9: Classification Report

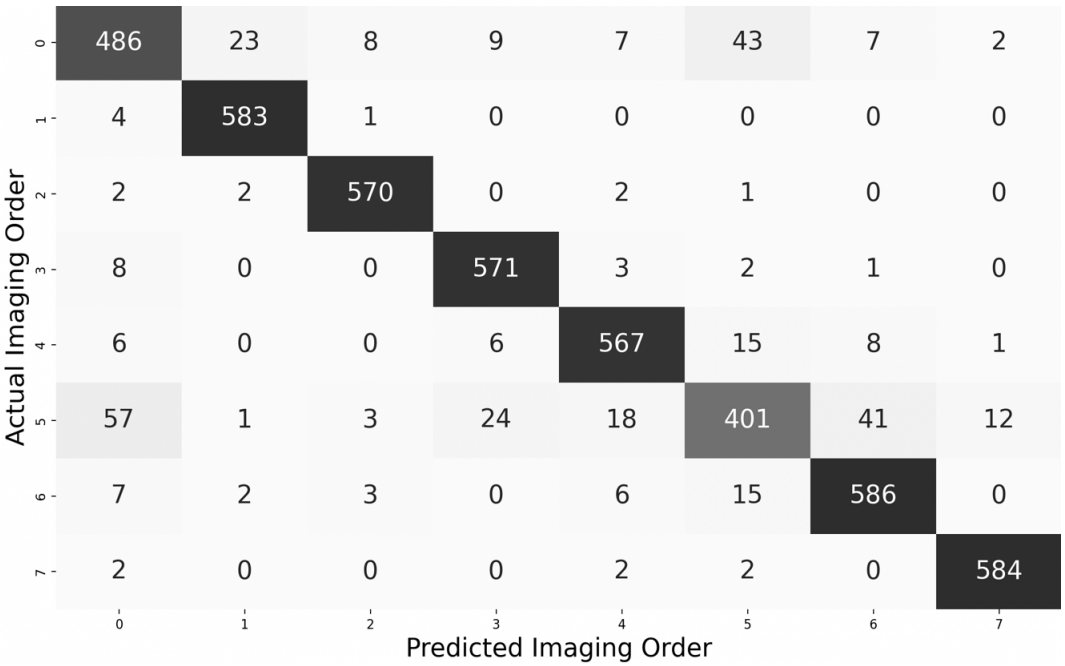

Figure 10: Heatmap

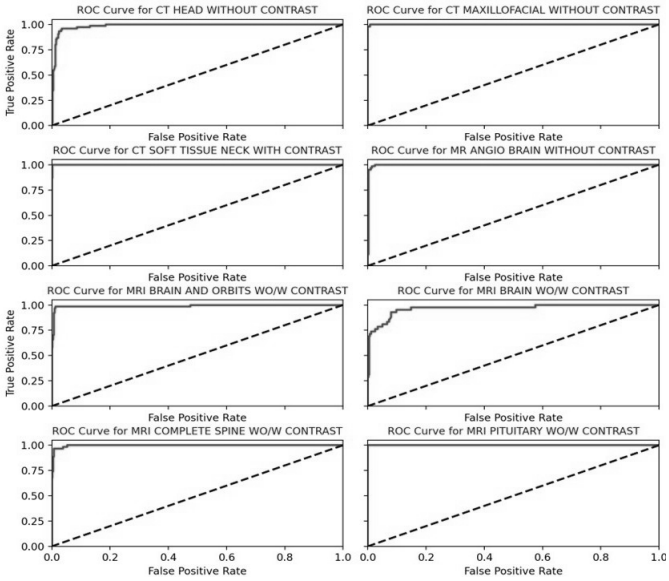

Figure 11: ROC Curves

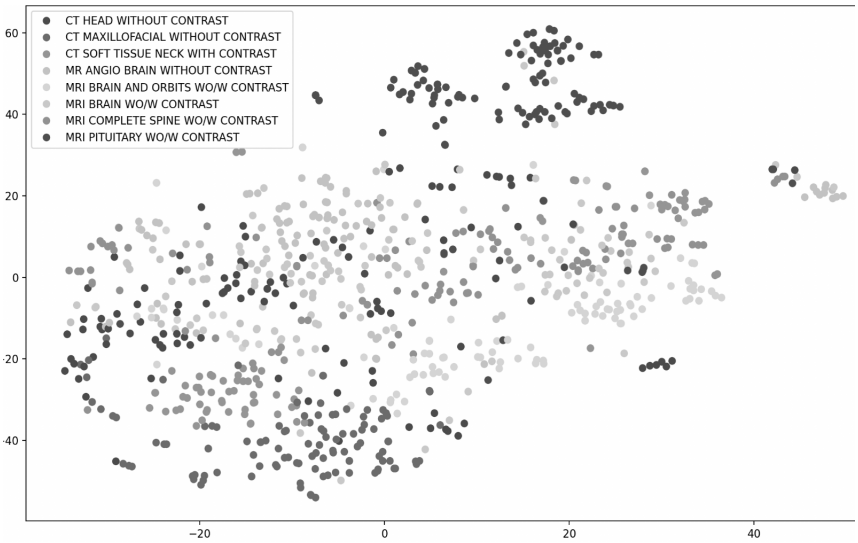

Figure 12: tSNE 2D

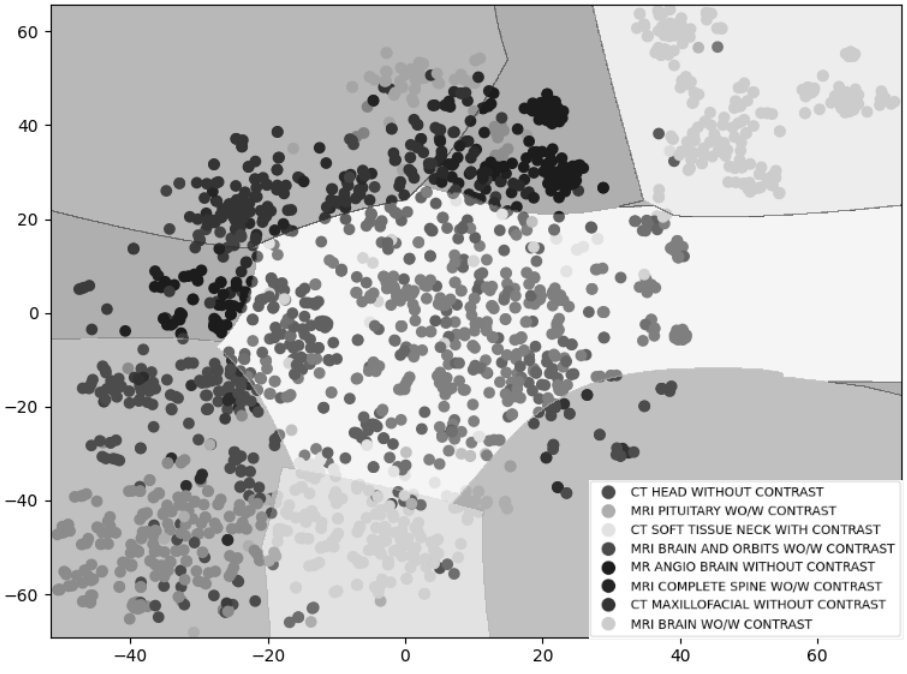

Figure 13: Decision Boundary Plot

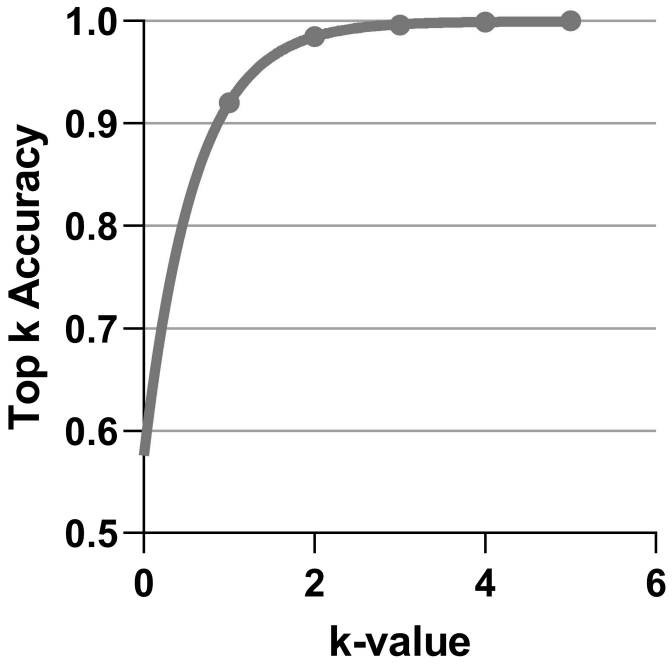

Figure 14: Top K Accuracy

Acknowledgements

The authors would like to thank Texas Children's Hospital for providing the project dataset and offering project feedback from a medical perspective.

Author Contributions

All authors contributed equally to the writing of the manuscript.

Conflicts of Interest

The authors have not declared any conflicts of interest.

References

1. Beek EJ van and Hoffman EA. Functional imaging: CT and MRI. Clinics in chest medicine 2008;29:195–216.
2. Weishaupt D, Koechli VD, Marincek B, and Kim EE. How does MRI Work? An introduction to the physics and function of magnetic resonance imaging. Journal of Nuclear Medicine 2007;48:1910–0.

3. Geva T. Magnetic resonance imaging: historical perspective. Journal of cardiovascular magnetic resonance 2006;8:573–80.

4. Plewes DB and Kucharczyk W. Physics of MRI: a primer. Journal of magnetic resonance imaging 2012;35:1038–54.

5. Gossuin Y, Hocq A, Gillis P, and Lam VQ. Physics of magnetic resonance imaging: from spin to pixel. Journal of Physics D: Applied Physics 2010;43:213001.

6. Pooley RA. Fundamental physics of MR imaging. Radiographics 2005;25:1087–99.

7. Stafford RJ. The physics of magnetic resonance imaging safety. Magnetic Resonance Imaging Clinics 2020;28:517–36.

8. McCollough CH, Bushberg JT, Fletcher JG, and Eckel LJ. Answers to common questions about the use and safety of CT scans. In: *Mayo Clinic Proceedings*. Vol. 90. 10. Elsevier. 2015:1380–92.

9. Lee CI, Haims AH, Monico EP, Brink JA, and Forman HP. Diagnostic CT scans: assessment of patient, physician, and radiologist awareness of radiation dose and possible risks. Radiology 2004;231:393–8.

10. Sluimer I, Schilham A, Prokop M, and Van Ginneken B. Computer analysis of computed tomography scans of the lung: a survey. IEEE transactions on medical imaging 2006;25:385–405.

11. Iglehart JK. Health insurers and medical-imaging policy—a work in progress. New England Journal of Medicine 2009;360:1030–7.

12. McCollough C and Leng S. Use of artificial intelligence in computed tomography dose optimisation. Annals of the ICRP 2020;49:113–25.

13. Lehnert BE and Bree RL. Analysis of appropriateness of outpatient CT and MRI referred from primary care clinics at an academic medical center: how critical is the need for improved decision support? Journal of the American College of Radiology 2010;7:192–7.

14. Pourjabbar S, Cavallo JJ, Arango J, et al. Impact of radiologist-driven change-order requests on outpatient CT and MRI examinations. Journal of the American College of Radiology 2020;17:1014–24.

15. Brenner DJ and Hall EJ. Computed tomography—an increasing source of radiation exposure. New England journal of medicine 2007;357:2277–84.

16. Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA pediatrics 2013;167:700–7.

17. Meulepas JM, Ronckers CM, Smets AM, et al. Radiation exposure from pediatric CT scans and subsequent cancer risk in the Netherlands. JNCI: Journal of the National Cancer Institute 2019;111:256–63.

18. Hall E and Brenner D. Cancer risks from diagnostic radiology. The British journal of radiology 2008;81:362–78.

19. Mason KP, Lubisch NB, Robinson F, and Roskos R. Intramuscular dexmedetomidine sedation for pediatric MRI and CT. American Journal of Roentgenology 2011;197:720–5.

20. Kamat PP, McCracken CE, Simon HK, et al. Trends in outpatient procedural sedation: 2007–2018. Pediatrics 2020;145.

21. Flaherty S, Zepeda ED, Mortele K, and Young GJ. Magnitude and financial implications of inappropriate diagnostic imaging for three common clinical conditions. International Journal for Quality in Health Care 2019;31:691–7.

22. Blackmore CC, Mecklenburg RS, and Kaplan GS. Effectiveness of clinical decision support in controlling inappropriate imaging. Journal of the American College of Radiology 2011;8:19–25.

23. Jebb AT, Parrigon S, and Woo SE. Exploratory data analysis as a foundation of inductive research. Human Resource Management Review 2017;27:265–76.

24. Mohammed R, Rawashdeh J, and Abdullah M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. 2020:243–8. DOI: 10.1109/ICICS49469.2020.239556.

25. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. CoRR 2020;abs/2005.14165.

26. Loper E and Bird S. NLTK: The Natural Language Toolkit. 2002. arXiv: cs/0205028 [cs.CL].

27. Khyani D and B S S. An Interpretation of Lemmatization and Stemming in Natural Language Processing. Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology 2021;22:350–7.

28. Jivani AG et al. A comparative study of stemming algorithms. Int. J. Comp. Tech. Appl 2011;2:1930–8.

29. Pramana RS, Debora, Subroto JJ, Gunawan AAS, and Anderies. Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity. 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA) 2022:1–6.

30. Balakrishnan V and Lloyd-Yemoh E. Stemming and lemmatization: A comparison of retrieval performances. 2014.

31. KS K and Sangeetha S. SECNLP: A Survey of Embeddings in Clinical Natural Language Processing. CoRR 2019;abs/1903.01039.

32. Guo C and Berkhahn F. Entity Embeddings of Categorical Variables. CoRR 2016;abs/1604.06737.

33. Chawda A, Grimm S, and Kloft M. Unsupervised Anomaly Detection for Auditing Data and Impact of Categorical Encodings. 2022.

34. SATHVIK M. Enhancing Machine Learning Algorithms using GPT Embeddings for Binary Classification. 2023.

35. Banerjee I, Chen M, Lungren M, and Rubin D. Radiology Report Annotation using Intelligent Word Embeddings: Applied to Multi-institutional Chest CT Cohort. Journal of Biomedical Informatics 2017;77.

36. Fernández A, Garcia S, Herrera F, and Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research 2018;61:863–905.

37. Blagus R and Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics 2013;14:1–16.

38. Mohammed R, Rawashdeh J, and Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. 2020:243–8.

39. Clarke R, Ressom HW, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer 2008;8:37–49.

40. Hussain H. Robustness of Support Vector Machines. Dissertation. 2020.

41. Pappu V and Pardalos P. High Dimensional Data Classification. In: 2013:34. DOI: 10.1007/978-1-4939-0742-7_8.

42. Ghaddar B and Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. European Journal of Operational Research 2018;265:993–1004.

43. Joshi M, Pakhomov S, Pedersen T, and Chute CG. A comparative study of supervised learning as applied to acronym expansion in clinical reports. In: *AMIA annual symposium proceedings*. Vol. 2006. American Medical Informatics Association. 2006:399.

44. Ahmad GN, Fatima H, Ullah S, Salah Saidi A, and Imdadullah. Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV. IEEE Access 2022;10:80151–73.

45. Deshwal V and Sharma M. Breast cancer detection using SVM classifier with grid search technique. International Journal of Computer Applications 2019;975:8887.

46. Fuadi AZ, Haq IN, Leksono E, et al. Support Vector Machine to Predict Electricity Consumption in the Energy Management Laboratory. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) 2021;5:466–73.

47. Goldberg Y and Elhadad M. splitSVM: Fast, Space-Efficient, non-Heuristic, Polynomial Kernel Computation for NLP Applications. In: *Proceedings of ACL-08: HLT, Short Papers*. Columbus, Ohio: Association for Computational Linguistics, 2008:237–40. URL: https://aclanthology.org/P08-2060.

48. Chang YW, Hsieh CJ, Chang KW, Ringgaard M, and Lin CJ. Training and testing low-degree polynomial data mappings via linear SVM. Journal of Machine Learning Research 2010;11.

49. Berge L. Diagnostic Imaging for the Emergency Physician. JAMA : the journal of the American Medical Association 2012;308:189.

50. Hussain S, Mubeen I, Ullah N, et al. Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. BioMed Research International 2022;2022:5164970.

51. Tith S, Lalwani K, and Fu R. Complications of three deep sedation methods for magnetic resonance imaging. Journal of Anaesthesiology Clinical Pharmacology 2012;28:178–84.

52. Jessome R. Improving patient flow in diagnostic imaging: a case report. J. Med. Imaging Radiat. Sci. 2020;51:678–88.

53. Sistrom CL and McKay NL. Costs, charges, and revenues for hospital diagnostic imaging procedures: differences by modality and hospital characteristics. J. Am. Coll. Radiol. 2005;2:511–9.

54. Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, and Forghani R. Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. Radiology: Artificial Intelligence 2023;5:e220028.