

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

Videet Mehta^{1†}, Rohan Dharia^{2†}, and Nilesh Desai^{3,4}

¹Math and Science Academy, Dulles High School, Sugar Land, Texas.

²Global Studies Academy, Travis High School, Richmond, Texas.

³Department 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.

32 **1 Introduction**

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

39 **1.1 Purpose**

40 Although their general purpose is the same, MRI and CT scans are necessary for different diagnosis
41 requirements. MRI is a non-radiation imaging process that utilizes the nuclear magnetic resonance
42 phenomenon to generate images of soft tissue structures. By aligning hydrogen atoms in the body
43 with the magnetic field and perturbing them with radio waves, the MRI scanner measures the result-
44 ing energy release and constructs a three-dimensional representation of the tissues being examined
45 [2–7]. In contrast, a CT scan conducts numerous X-ray projections to create detailed cross-sectional
46 images of the body with millimeter precision[8–10]. An intramedullary tumor, for example, would
47 require an MRI scan due to its sensitivity in detecting soft tissue abnormalities. A broken bone,
48 however, would require a CT scan due to its ability to visualize dense structures.

49 **1.2 Imaging**

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

60 **1.3 Concerns**

61 Imaging appropriateness has been of important concern in the effort to improve medical imaging
62 efficiency and cost reduction. A past study showed that 26% of medical imaging orders were identified
63 as inappropriate [13, 14]. This high proportion has four main consequences for patients. (i) Re-
64 imaging leads to excessive radiation, which has a well-established link to cancer and other disorders
65 [15–18]. (ii) MRI/CT scans require pediatric patients to be sedated, and repeating this can cause
66 cardiopulmonary complications [19, 20]. (iii) The patient recovery timeline can be delayed due
67 to the aforementioned long process of imaging having to be repeated, and for a patient with an

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

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

81 **2 Methods**

82 **2.1 Data Collection**

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

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

92 The data in Figure 1 included information on various unidentifiable patient attributes namely
93 age, weight, and height. which were assessed for their contribution to the model’s output. It was
94 determined that these columns would introduce unnecessary features with low predictive power
95 due to their randomness and were therefore excluded. The average word count of the clinical
96 signs/symptoms showed an average of 4.6 words indicating the limited information provided with
97 each entry (Fig. 3). Three key issues were identified in the data that impacted our approach prior
98 to algorithm implementation. First, the degree of irregularity present required the implementation
99 of a pre-processing step prior to ML. Many rows consisted of abbreviations, irregular capitalization,
100 incorrect spelling and grammar, and empty entries. Second, a severe data imbalance, with nearly
101 11,000 cases of CT Head without Contrast and just 1,000 of MRI Pituitary indicated the need
102 for class balancing, either with under-sampling or oversampling [24]. Lastly, the most frequent
103 word count of the symptom text rows was found to be just three, as seen in the distribution graph,
104 indicating that each word present contributes significant meaning and that any pre-processing should
105 be somewhat conservative (Fig. 3).

106 2.2 Preprocessing

107 A three-step processing algorithm addressed these data irregularities. The initial step was to nor-
108 malize and condense the symptom text. Our initial approach was to utilize OpenAI’s Davinci GPT-3
109 API [25] to ask it to extract important keywords and include similar words to the existing prompt
110 for additional data. This approach proved to be costly and computationally expensive as it required
111 API and tokens for purchase. The alternative was the Natural Language Toolkit (NLTK) which
112 tokenizes each symptom row and filters through the StopWord corpus to identify stop-words [26].
113 As opposed to extracting important keywords, the algorithm uses the corpus to identify and extract
114 unnecessary filler words that serve no significant meaning such as “with”, “to”, and “also”. It proved
115 to be the most effective approach in normalizing the data and accounted for special characters and
116 irregular syntax as seen in Figure 5. However, using the NLTK library sacrificed the ability to
117 correct spellings, a feature that the GPT model would automatically account for.

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

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

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

146 the two existing points. SMOTE has also been proven to work well in high-dimensional datasets
147 [37]. Each of the minority classes was oversampled to match the count of the majority class (CT
148 Head without Contrast). SMOTE increased the sample size from 40,667 to approximately 90,000
149 rows and was used to train the final ML model (Fig. 7). Basic minority oversampling and majority
150 undersampling techniques were also tested, however, the SMOTE algorithm was found to be more
151 effective at increasing the variety of data present in a limited dataset [38]. By generating synthetic
152 data points, the algorithm was able to maintain the underlying structure of the data while eliminating
153 the problem of class imbalance.

154 2.3 Machine Learning

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

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

$$172 \beta_0 * x_1 + \beta_2 * x_2 = -\beta_0 \quad (1)$$

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

175 Equation 1 can be generalized to:

$$176 \beta_0 + \beta_0 * x_1 + \beta_1 * x_2 + \dots + \beta_p * x_{p+1} = 0 \quad (2)$$

177 or

$$178 \vec{w} \bullet \vec{x} + b = 0 \quad (3)$$

179 X_i is defined as the vector to display all features of index i . With \vec{X}_+ describing the positive class
180 and \vec{X}_- describing the negative class. The support vectors can be defined as

$$181 \vec{w} \bullet \vec{X}_+ + b > 1$$

182 and

$$183 \vec{w} \bullet \vec{X}_- + b < -1 \quad (4)$$

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

$$185 \quad Y_i(\vec{w} \bullet \vec{X}_i + b) \geq 1 \quad (5)$$

186 The width of the margin is then defined as

$$187 \quad \frac{(\vec{X}_+ - \vec{X}_-) \bullet \vec{w}}{\|\vec{w}\|} \quad (6)$$

$$188 \quad \text{or} \\ 189 \quad \frac{2}{\|\vec{w}\|} \quad (7)$$

190 And to maximize Equation 7, the model must minimize

$$191 \quad \frac{1}{2} \|\vec{w}\|^2 \quad (8)$$

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

$$193 \quad L = \frac{1}{2} \|\vec{w}\|^2 - \sum \alpha_i [Y_i(\vec{w} \bullet \vec{X}_i + b) - 1] \quad (9)$$

194 Since

$$195 \quad \frac{\delta L}{\delta \vec{w}} = 0, \text{ therefore } \vec{w} = \sum \alpha_i Y_i X_i \quad (10)$$

196 and

$$197 \quad \frac{\delta L}{\delta b} = 0, \text{ therefore } \sum X_i Y_i = 0 \quad (11)$$

198 Which leads to

$$199 \quad L = -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j Y_i Y_j \vec{X}_i \bullet \vec{X}_j + \sum \alpha_i \quad (12)$$

200 From here, we see that the SVC depends on $\vec{X}_i \bullet \vec{X}_j$.

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

$$204 \quad Accuracy = \frac{TP+FP}{TP+FP+FN+TN} \quad (13)$$

205 After selecting a model, the support vector machine further optimized the model's performance
206 through GridSearchCV which further optimized the model's performance by conducting tuning of
207 six model parameters using GridSearchCV. The GridSearchCV starts with random hyperparameters
208 and slowly optimized the model through a trial-and-error method. The perfect model is met when
209 there is a balance between underfitting and overfitting, therefore increasing the metrics of the model
210 [44–46]. The parameters that were optimized were Kernel, C, Gamma, and Train-Test split. A
211 low-degree polynomial kernel is optimal in NLP situations [47, 48].

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

$$214 \quad K(x, y) = (x^T y + c)^d = (\varphi(x), \varphi(y)) \quad (14)$$

215 3 Results

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

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

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

$$225 \textit{Precision} = \frac{TP}{TP+FP} (15)$$

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

$$228 \textit{Recall} = \frac{TP}{TP+FN} (16)$$

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

$$230 \textit{F1} = \frac{\textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}} (17)$$

231 MRI Pituitary WO/W Contrast had the highest scores of all three metrics at 0.97, 0.99, and
232 0.98 respectively, suggesting confidence with this class. MRI Brain WO/W Contrast had the lowest
233 scores of all three metrics at 0.84, 0.72, and 0.77 respectively, suggesting that this class may be
234 highly similar to another. However, the average of all individual class metrics came out to be 0.92,
235 0.92, and 0.92 respectively, confirming the model's overall accuracy.

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

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

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

$$251 p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} (18)$$

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

$$253 \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \quad (19)$$

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

$$257 \quad q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_k - y_i\|^2)^{-1}} \quad (20)$$

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

$$261 \quad D_{KL}(P \parallel Q) = \sum_{x \in \chi} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \quad (21)$$

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

$$266 \quad C = KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right) \quad (22)$$

267 The results are shown in Fig. 12.

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

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

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

284 4 Discussion

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

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

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

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

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

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

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

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

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

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

332 **5 Conclusion**

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

339 **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

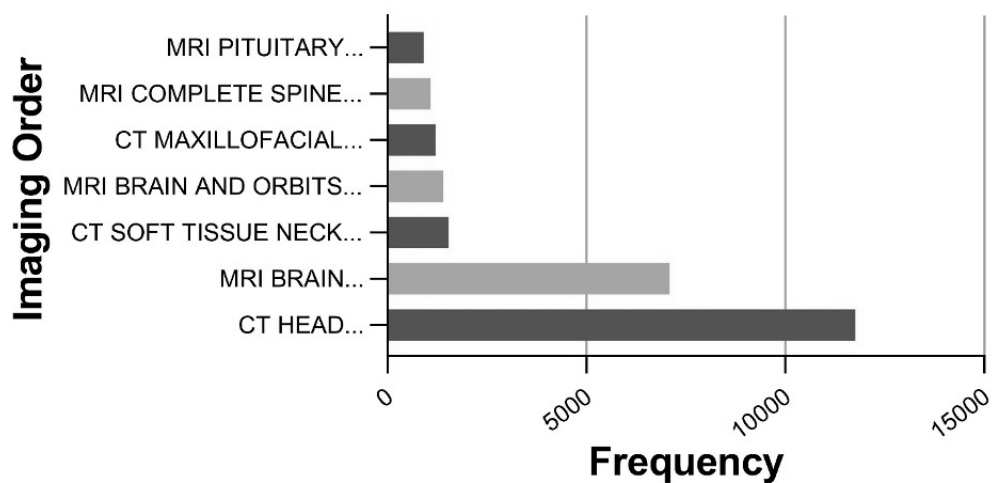


Figure 2: Class Distribution

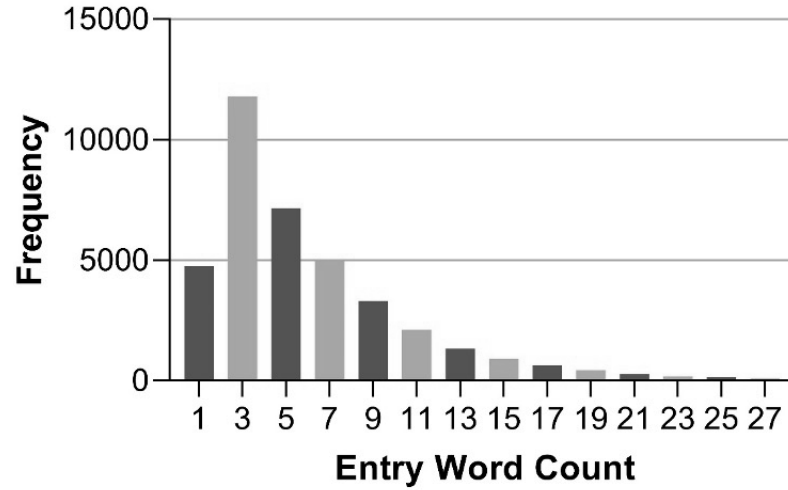


Figure 3: Word Count Distribution

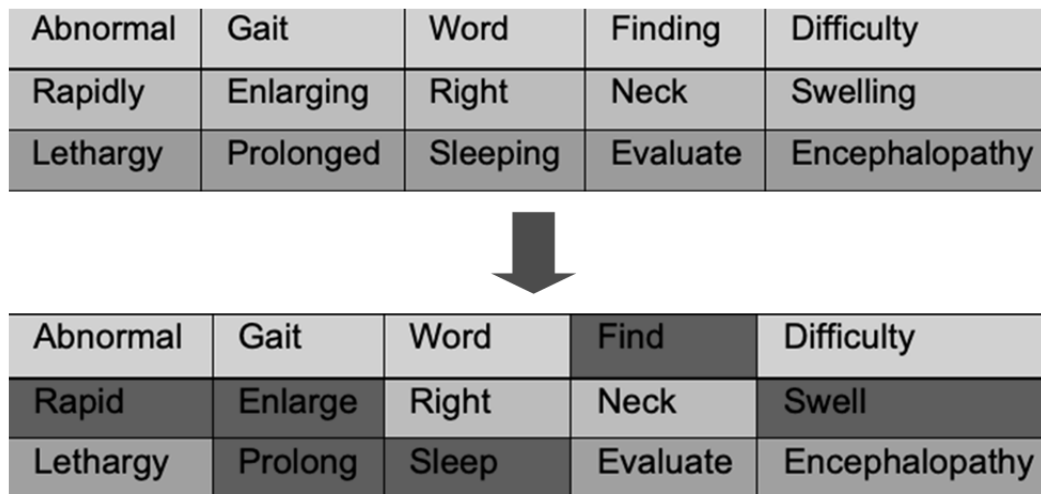


Figure 4: Lemmatization

Recurrent	Focal	Seizures	With	Eye	Deviation
Dog	Bite	To	Left	Face	Eval
Female	With	No	Berlin	Continued	Concern

↓

Recurrent	Focal	Seizures		Eye	Deviation
Dog	Bite		Left	Face	Eval
Female			Berlin	Continued	Concern

Figure 5: Keyword Filtering

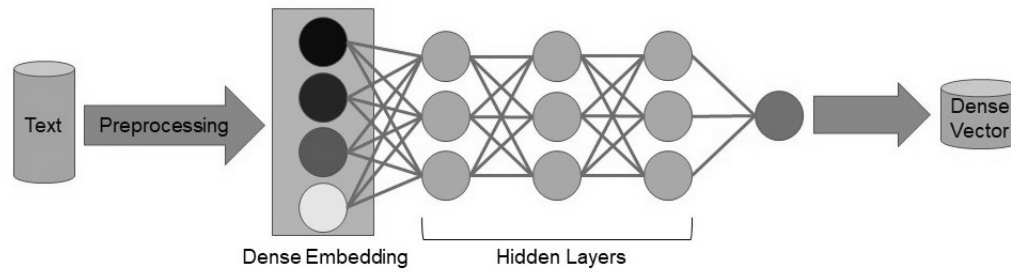


Figure 6: Entity Embedding

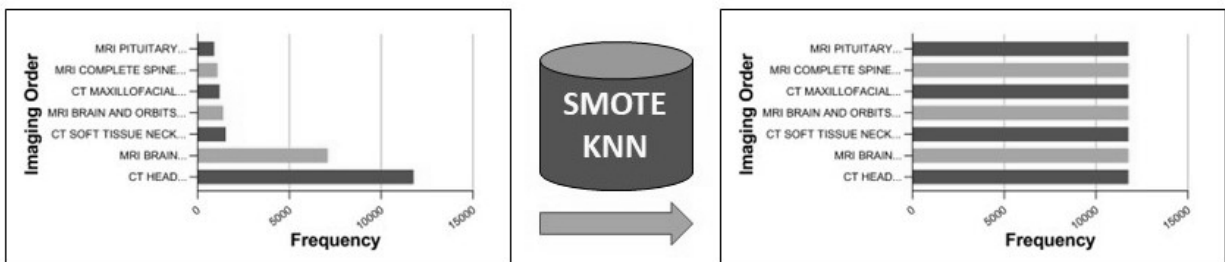


Figure 7: Synthetic Sampling

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 WOW/CONTRAST	603	0.94	0.94	0.94
MRI BRAIN WOW/CONTRAST	557	0.84	0.72	0.77
MRI COMPLETE SPINE WOW/CONTRAST	617	0.91	0.95	0.93
MRI PITUITARY WOW/CONTRAST	590	0.97	0.99	0.98
Sum/Average	4704	0.92	0.92	0.92

Figure 9: Classification Report

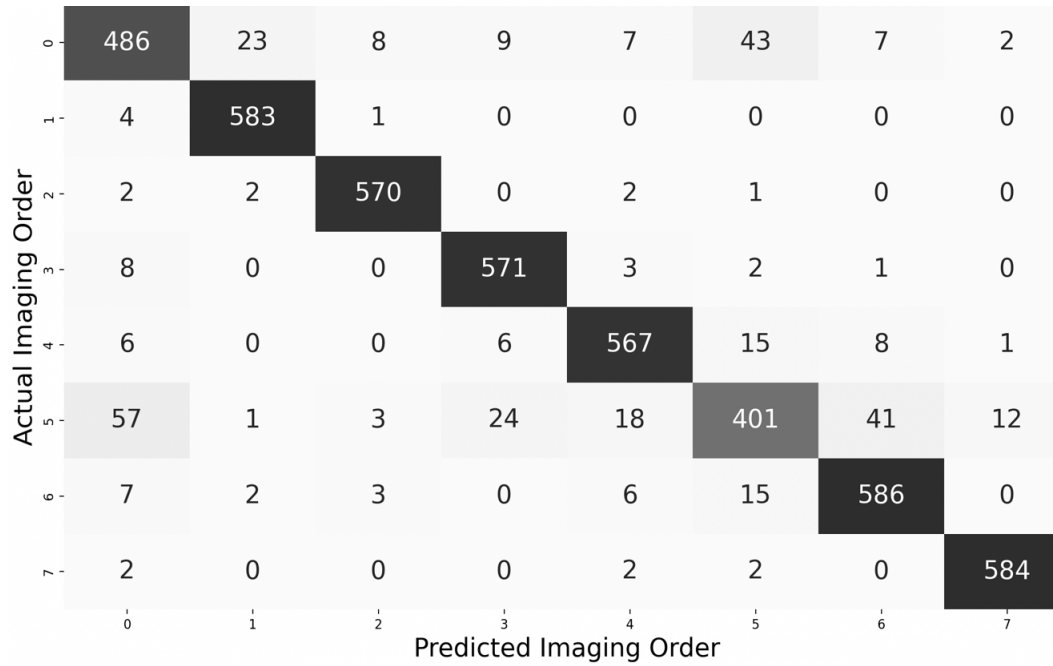


Figure 10: Heatmap

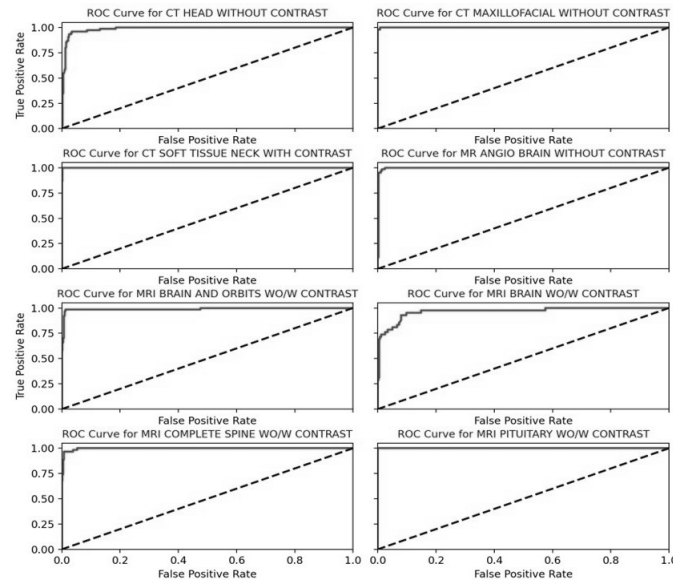


Figure 11: ROC Curves

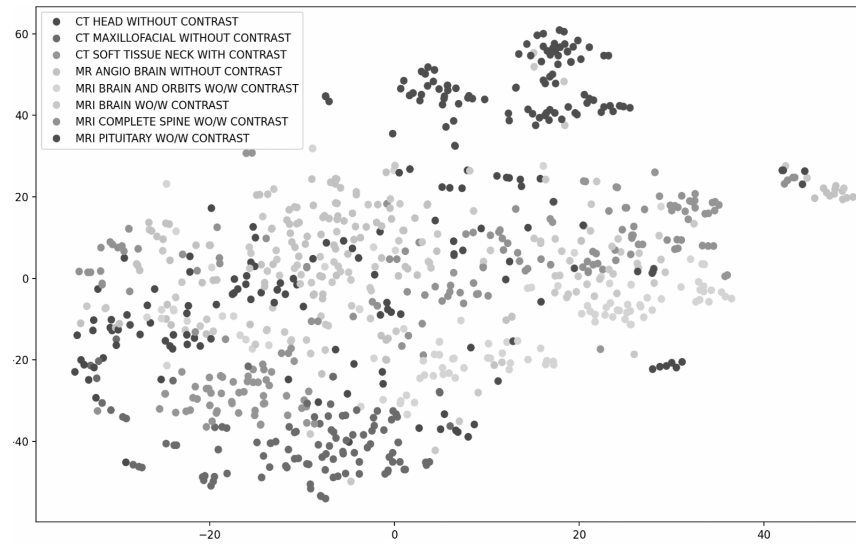


Figure 12: tSNE 2D

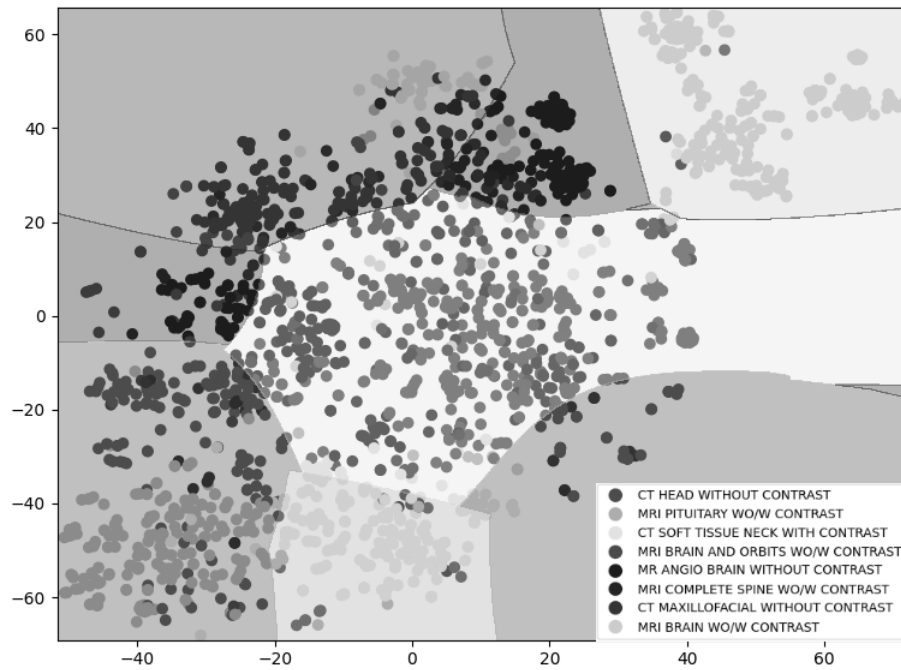


Figure 13: Decision Boundary Plot

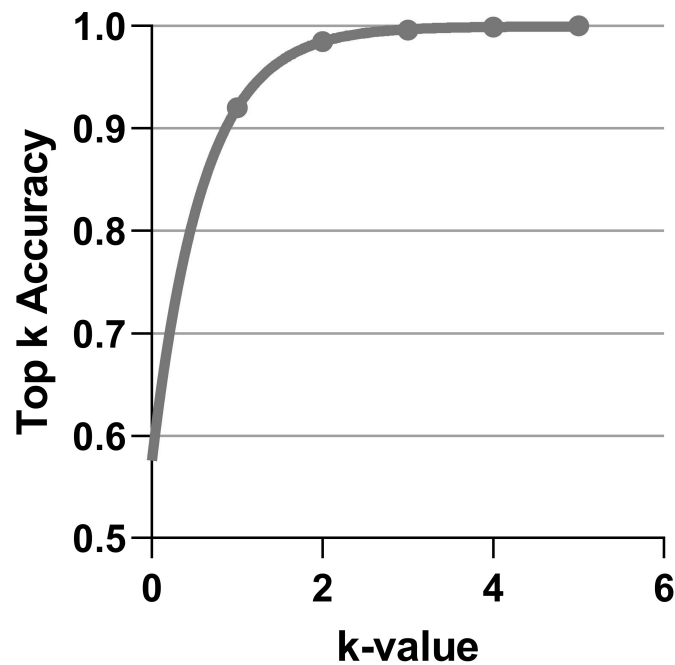


Figure 14: Top K Accuracy

340 Acknowledgements

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

343 Author Contributions

344 All authors contributed equally to the writing of the manuscript.

345 Conflicts of Interest

346 The authors have not declared any conflicts of interest.

347 References

- 348 1. Beek EJ van and Hoffman EA. Functional imaging: CT and MRI. Clinics in chest medicine
349 2008;29:195–216.
- 350 2. Weishaupt D, Koechli VD, Marincek B, and Kim EE. How does MRI Work? An introduc-
351 tion to the physics and function of magnetic resonance imaging. Journal of Nuclear Medicine
352 2007;48:1910–0.

- 353 3. Geva T. Magnetic resonance imaging: historical perspective. *Journal of cardiovascular magnetic*
354 *resonance* 2006;8:573–80.
- 355 4. Plewes DB and Kucharczyk W. Physics of MRI: a primer. *Journal of magnetic resonance*
356 *imaging* 2012;35:1038–54.
- 357 5. Gossuin Y, Hocq A, Gillis P, and Lam VQ. Physics of magnetic resonance imaging: from spin
358 to pixel. *Journal of Physics D: Applied Physics* 2010;43:213001.
- 359 6. Pooley RA. Fundamental physics of MR imaging. *Radiographics* 2005;25:1087–99.
- 360 7. Stafford RJ. The physics of magnetic resonance imaging safety. *Magnetic Resonance Imaging*
361 *Clinics* 2020;28:517–36.
- 362 8. McCollough CH, Bushberg JT, Fletcher JG, and Eckel LJ. Answers to common questions about
363 the use and safety of CT scans. In: *Mayo Clinic Proceedings*. Vol. 90. 10. Elsevier. 2015:1380–92.
- 364 9. Lee CI, Haims AH, Monico EP, Brink JA, and Forman HP. Diagnostic CT scans: assessment
365 of patient, physician, and radiologist awareness of radiation dose and possible risks. *Radiology*
366 2004;231:393–8.
- 367 10. Sluimer I, Schilham A, Prokop M, and Van Ginneken B. Computer analysis of computed
368 tomography scans of the lung: a survey. *IEEE transactions on medical imaging* 2006;25:385–
369 405.
- 370 11. Iglehart JK. Health insurers and medical-imaging policy—a work in progress. *New England*
371 *Journal of Medicine* 2009;360:1030–7.
- 372 12. McCollough C and Leng S. Use of artificial intelligence in computed tomography dose optimi-
373 sation. *Annals of the ICRP* 2020;49:113–25.
- 374 13. Lehnert BE and Bree RL. Analysis of appropriateness of outpatient CT and MRI referred
375 from primary care clinics at an academic medical center: how critical is the need for improved
376 decision support? *Journal of the American College of Radiology* 2010;7:192–7.
- 377 14. Pourjabbar S, Cavallo JJ, Arango J, et al. Impact of radiologist-driven change-order requests
378 on outpatient CT and MRI examinations. *Journal of the American College of Radiology*
379 2020;17:1014–24.
- 380 15. Brenner DJ and Hall EJ. Computed tomography—an increasing source of radiation exposure.
381 *New England journal of medicine* 2007;357:2277–84.
- 382 16. Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pedi-
383 atrics and the associated radiation exposure and estimated cancer risk. *JAMA pediatrics*
384 2013;167:700–7.
- 385 17. Meulepas JM, Ronckers CM, Smets AM, et al. Radiation exposure from pediatric CT scans
386 and subsequent cancer risk in the Netherlands. *JNCI: Journal of the National Cancer Institute*
387 2019;111:256–63.
- 388 18. Hall E and Brenner D. Cancer risks from diagnostic radiology. *The British journal of radiology*
389 2008;81:362–78.

- 390 19. Mason KP, Lubisch NB, Robinson F, and Roskos R. Intramuscular dexmedetomidine sedation
391 for pediatric MRI and CT. *American Journal of Roentgenology* 2011;197:720–5.
- 392 20. Kamat PP, McCracken CE, Simon HK, et al. Trends in outpatient procedural sedation: 2007–
393 2018. *Pediatrics* 2020;145.
- 394 21. Flaherty S, Zepeda ED, Morteale K, and Young GJ. Magnitude and financial implications of
395 inappropriate diagnostic imaging for three common clinical conditions. *International Journal
396 for Quality in Health Care* 2019;31:691–7.
- 397 22. Blackmore CC, Mecklenburg RS, and Kaplan GS. Effectiveness of clinical decision support in
398 controlling inappropriate imaging. *Journal of the American College of Radiology* 2011;8:19–25.
- 399 23. Jebb AT, Parrigon S, and Woo SE. Exploratory data analysis as a foundation of inductive
400 research. *Human Resource Management Review* 2017;27:265–76.
- 401 24. Mohammed R, Rawashdeh J, and Abdullah M. Machine Learning with Oversampling and
402 Undersampling Techniques: Overview Study and Experimental Results. In: *2020 11th Inter-
403 national Conference on Information and Communication Systems (ICICS)*. 2020:243–8. DOI:
404 10.1109/ICICS49469.2020.239556.
- 405 25. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. *CoRR* 2020;abs/2005.14165.
- 406 26. Loper E and Bird S. NLTK: The Natural Language Toolkit. 2002. arXiv: [cs/0205028](https://arxiv.org/abs/cs/0205028) [cs.CL].
- 407 27. Khyani D and B S S. An Interpretation of Lemmatization and Stemming in Natural Language
408 Processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and
409 Technology* 2021;22:350–7.
- 410 28. Jivani AG et al. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*
411 2011;2:1930–8.
- 412 29. Pramana RS, Debora, Subroto JJ, Gunawan AAS, and Anderies. Systematic Literature Review
413 of Stemming and Lemmatization Performance for Sentence Similarity. 2022 IEEE 7th Inter-
414 national Conference on Information Technology and Digital Applications (ICITDA) 2022:1–
415 6.
- 416 30. Balakrishnan V and Lloyd-Yemoh E. Stemming and lemmatization: A comparison of retrieval
417 performances. 2014.
- 418 31. KS K and Sangeetha S. SECNLP: A Survey of Embeddings in Clinical Natural Language
419 Processing. *CoRR* 2019;abs/1903.01039.
- 420 32. Guo C and Berkhahn F. Entity Embeddings of Categorical Variables. *CoRR* 2016;abs/1604.06737.
- 421 33. Chawda A, Grimm S, and Kloft M. Unsupervised Anomaly Detection for Auditing Data and
422 Impact of Categorical Encodings. 2022.
- 423 34. SATHVIK M. Enhancing Machine Learning Algorithms using GPT Embeddings for Binary
424 Classification. 2023.
- 425 35. Banerjee I, Chen M, Lungren M, and Rubin D. Radiology Report Annotation using Intelligent
426 Word Embeddings: Applied to Multi-institutional Chest CT Cohort. *Journal of Biomedical
427 Informatics* 2017;77.

- 428 36. Fernández A, Garcia S, Herrera F, and Chawla NV. SMOTE for learning from imbalanced
429 data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence*
430 *research* 2018;61:863–905.
- 431 37. Blagus R and Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics*
432 2013;14:1–16.
- 433 38. Mohammed R, Rawashdeh J, and Abdullah M. Machine learning with oversampling and un-
434 dersampling techniques: overview study and experimental results. 2020:243–8.
- 435 39. Clarke R, Resson HW, Wang A, et al. The properties of high-dimensional data spaces: im-
436 plications for exploring gene and protein expression data. *Nature Reviews Cancer* 2008;8:37–
437 49.
- 438 40. Hussain H. Robustness of Support Vector Machines. Dissertation. 2020.
- 439 41. Pappu V and Pardalos P. High Dimensional Data Classification. In: 2013:34. DOI: 10.1007/
440 978-1-4939-0742-7_8.
- 441 42. Ghaddar B and Naoum-Sawaya J. High dimensional data classification and feature selection
442 using support vector machines. *European Journal of Operational Research* 2018;265:993–1004.
- 443 43. Joshi M, Pakhomov S, Pedersen T, and Chute CG. A comparative study of supervised learning
444 as applied to acronym expansion in clinical reports. In: *AMIA annual symposium proceedings*.
445 Vol. 2006. American Medical Informatics Association. 2006:399.
- 446 44. Ahmad GN, Fatima H, Ullah S, Salah Saidi A, and Imdadullah. Efficient Medical Diagnosis of
447 Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV.
448 *IEEE Access* 2022;10:80151–73.
- 449 45. Deshwal V and Sharma M. Breast cancer detection using SVM classifier with grid search
450 technique. *International Journal of Computer Applications* 2019;975:8887.
- 451 46. Fuadi AZ, Haq IN, Leksono E, et al. Support Vector Machine to Predict Electricity Consump-
452 tion in the Energy Management Laboratory. *Jurnal RESTI (Rekayasa Sistem dan Teknologi*
453 *Informasi)* 2021;5:466–73.
- 454 47. Goldberg Y and Elhadad M. splitSVM: Fast, Space-Efficient, non-Heuristic, Polynomial Kernel
455 Computation for NLP Applications. In: *Proceedings of ACL-08: HLT, Short Papers*. Columbus,
456 Ohio: Association for Computational Linguistics, 2008:237–40. URL: <https://aclanthology.org/P08-2060>.
- 457
- 458 48. Chang YW, Hsieh CJ, Chang KW, Ringgaard M, and Lin CJ. Training and testing low-degree
459 polynomial data mappings via linear SVM. *Journal of Machine Learning Research* 2010;11.
- 460 49. Berge L. Diagnostic Imaging for the Emergency Physician. *JAMA : the journal of the American*
461 *Medical Association* 2012;308:189.
- 462 50. Hussain S, Mubeen I, Ullah N, et al. Modern Diagnostic Imaging Technique Applications and
463 Risk Factors in the Medical Field: A Review. *BioMed Research International* 2022;2022:5164970.
- 464 51. Tith S, Lalwani K, and Fu R. Complications of three deep sedation methods for magnetic
465 resonance imaging. *Journal of Anaesthesiology Clinical Pharmacology* 2012;28:178–84.

- 466 52. Jessome R. Improving patient flow in diagnostic imaging: a case report. *J. Med. Imaging Radiat.*
467 *Sci.* 2020;51:678–88.
- 468 53. Siström CL and McKay NL. Costs, charges, and revenues for hospital diagnostic imaging pro-
469 cedures: differences by modality and hospital characteristics. *J. Am. Coll. Radiol.* 2005;2:511–
470 9.
- 471 54. Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, and Forghani R. Generalizability of Ma-
472 chine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. *Radiology:*
473 *Artificial Intelligence* 2023;5:e220028.