BMJ Health &
Care Informatics

Advancing equity in breast cancer care: natural language processing for analysing treatment outcomes in under-represented populations

Jung In Park ,[1] Jong Won Park,[2] Kexin Zhang,[3] Doyop Kim[4]

To cite: Park JI, Park JW, Zhang K, *et al*. Advancing equity in breast cancer care: natural language processing for analysing treatment outcomes in under-represented populations. *BMJ Health Care Inform* 2024;**31**:e100966. doi:10.1136/bmjhci-2023-100966

Received 16 November 2023
Accepted 21 June 2024

[1]University of California Irvine, Irvine, California, USA
[2]Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, South Korea
[3]Donald Bren School of Information & Computer Sciences, University of California Irvine, Irvine, California, USA
[4]Independent Researcher, Irvine, California, USA

Correspondence to
Dr Jung In Park;
junginp@uci.edu

ABSTRACT

Objective The study aimed to develop natural language processing (NLP) algorithms to automate extracting patient-centred breast cancer treatment outcomes from clinical notes in electronic health records (EHRs), particularly for women from under-represented populations.

Methods The study used clinical notes from 2010 to 2021 from a tertiary hospital in the USA. The notes were processed through various NLP techniques, including vectorisation methods (term frequency-inverse document frequency (TF-IDF), Word2Vec, Doc2Vec) and classification models (support vector classification, K-nearest neighbours (KNN), random forest (RF)). Feature selection and optimisation through random search and fivefold cross-validation were also conducted.

Results The study annotated 100 out of 1000 clinical notes, using 970 notes to build the text corpus. TF-IDF and Doc2Vec combined with RF showed the highest performance, while Word2Vec was less effective. RF classifier demonstrated the best performance, although with lower recall rates, suggesting more false negatives. KNN showed lower recall due to its sensitivity to data noise.

Discussion The study highlights the significance of using NLP in analysing clinical notes to understand breast cancer treatment outcomes in under-represented populations. The TF-IDF and Doc2Vec models were more effective in capturing relevant information than Word2Vec. The study observed lower recall rates in RF models, attributed to the dataset's imbalanced nature and the complexity of clinical notes.

Conclusion The study developed high-performing NLP pipeline to capture treatment outcomes for breast cancer in under-represented populations, demonstrating the importance of document-level vectorisation and ensemble methods in clinical notes analysis. The findings provide insights for more equitable healthcare strategies and show the potential for broader NLP applications in clinical settings.

INTRODUCTION

Breast cancer is the second leading cause of cancer deaths in US women, comprising 30% of new female cancer diagnoses.[1] It is the most common cancer across all ethnic

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Before this study, it was understood that breast cancer is the most prevalent cancer affecting women of all ethnic groups in the USA, with disparities in outcomes among different racial and ethnic groups.

⇒ The widespread use of electronic health records and advances in natural language processing (NLP) offered avenues for improved patient care through detailed data analysis; however, there was a gap in automated, detailed analysis of clinical notes, especially for breast cancer treatment outcomes in women from under-represented populations, necessitating this study.

WHAT THIS STUDY ADDS

⇒ This study contributes by developing a robust NLP pipeline to analyse clinical notes for breast cancer treatment outcomes in under-represented populations.

⇒ It demonstrates the effectiveness of specific text vectorisation methods (term frequency-inverse document frequency and Doc2Vec) combined with classification models, particularly random forest (RF), in extracting relevant treatment outcome data from clinical notes.

⇒ The study also reveals the challenges in achieving high recall rates in predictive models, highlighting the complexity of clinical data and the need for specialised NLP approaches.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study has significant implications for future research, clinical practice and health policy.

⇒ It underscores the potential of NLP in enhancing the understanding of breast cancer treatment outcomes, particularly for under-represented groups, thereby guiding more personalised and equitable healthcare strategies.

⇒ The findings could influence policy decisions related to healthcare data management and the integration of NLP techniques in clinical settings.

⇒ Moreover, the developed pipeline can be adapted for other clinical NLP applications, potentially broadening its impact beyond breast cancer research.

groups in the USA, but disparities exist in outcomes.[2] While white women have higher incidence rates, black and Hispanic women face higher mortality rates.[3 4] Additionally, the incidence is increasing rapidly among Asian/Pacific Islanders and American Indian/Alaska Natives.[4]

The widespread adoption of electronic health records (EHRs) offers promising opportunities for predicting future events using large amounts of data.[5] Especially, unstructured clinical notes contain important information often not captured in structured, coded formats.[6] For example, patient-reported outcomes from patients with cancer are often not captured in structured EHRs, but is increasingly found in unstructured or semi-structured text formats within EHRs, facilitating translational research and personalised care.[7–9] One common approach in clinical text analysis involves using a rule-based natural language processing (NLP) algorithm that leverages distinct medical keywords from clinical texts.[10 11] Specifically, with the advancements in neural language modelling, integrating neural networks with features extracted from this rule-based NLP method can be achieved by using word embedding models for feature extraction.[12] This approach allows for building a fully neural network-based pipeline that combines embedding models with supervised learning algorithms.[13]

In cancer research, incorporating clinical notes into analyses is crucial for capturing information on comprehensive symptoms and side effects that patients experience,[14] as it can provide insights into monitoring and individualised symptom management. Several studies have investigated breast cancer treatment outcomes using clinical notes and NLP[14–16]; however, research that specifically aims the capture of treatment side effects and patient-reported outcomes in patients with breast cancer from under-represented populations remains sparse. Addressing this research gap is important, because these populations face unique health disparities that impact treatment outcomes and patient care. Understanding these specific challenges and barriers enables the development of targeted interventions to mitigate disparities and enhance health outcomes. There is a clear need for an automated tool to capture symptoms and side effects from clinical notes, enabling accurate symptom management and tailored nursing care planning for those patients from under-represented populations.

The goal of this study was to develop NLP algorithms to automate the knowledge extraction process for patient-centred breast cancer treatment outcomes from clinical notes, aiming to gain valuable insights to improve care for those from under-represented populations. Specifically, we aimed to compare the effectiveness of these algorithms in providing scientific evidence for their use in the care of patients with breast cancer from under-represented populations.

METHODS

To harness the full potential of large health datasets from the EHRs and unique application of NLP techniques, we sourced EHR clinical notes dated 1 January 2010 to 31 August 2021 at a tertiary hospital in the USA, selecting patients who met the following criteria: (1) women from under-represented populations (Hispanic, American Indian or Alaska Native, Asian, black or African-American, Native Hawaiian or Other Pacific Islander or multiple race); (2) aged 18 years or greater; (3) diagnosed with invasive breast cancer; (4) had at least one follow-up visit at the medical centre after breast cancer treatment (ie, surgery, radiation therapy, chemotherapy, endocrine therapy or hormone therapy). We excluded the patients who were not followed up at the medical centre.

Overview of the NLP pipeline

In this study, we developed a classification model to predict a binary outcome: whether a side effect was observed in relation to breast cancer treatment, based on the text within a clinical note. Our approach involved a multistep process, as illustrated in figure 1. The process began with raw clinical notes from which text was extracted to train and test the downstream models. The extracted texts underwent preprocessing to ensure they were clean and normalised. Following preprocessing, the cleaned text corpus was used for text vectorisation. Additionally, we randomly sampled notes and had them annotated by clinical experts. After annotation, the texts were mapped into a feature vector space (vectorisation). We then selected the most impactful features and reduced the feature dimension (feature selection) to train a conventional classifier and predict the outcome using this feature vector. Subsequent sections provide a detailed description of each step involved.

Data preprocessing and annotation

To prepare text data for the NLP process, it must undergo preprocessing. This involves standard NLP cleaning techniques such as removing numbers, special characters and duplicated words; performing word tokenisation; removing stop words and applying stemming.[17] Once cleaned, these text data serve as a corpus to train a vectorisation model that converts input text into numerical form (feature vector). This vectorisation can proceed without explicit document annotation, relying on the text corpus of the clinical notes. In contrast, expert annotations are crucial for the classification phase, making it a supervised learning task. Notes were labelled as positive if they referenced side effects or symptoms of breast cancer treatment, adhering to guidelines from the American Cancer Society and American Society of Clinical Oncology.[18] A clinical expert annotated 100 notes, which were randomly selected from the original texts. Subsequently, the annotated data were divided into training and test sets using a 7:3 ratio.

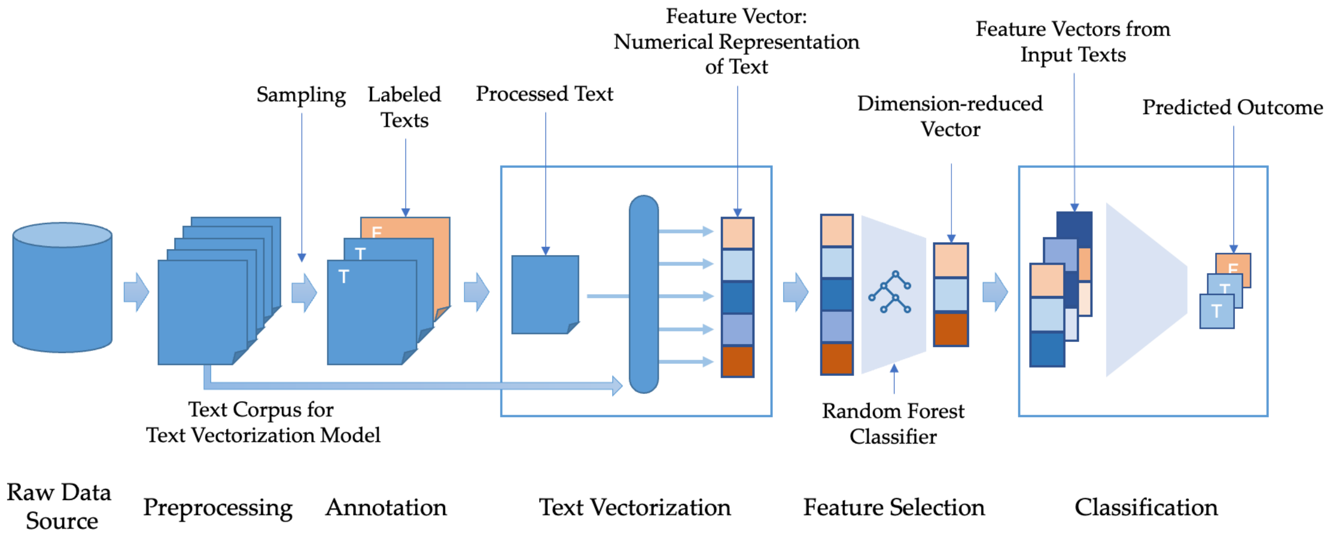

Figure 1 Overview of the natural language processing pipeline. T, true label; F, false label of clinical notes.

Text vectorisation

The texts were converted into a set of numerical values—a vector that represents a given text. We used three different vectorisation approaches—term frequency-inverse document frequency (TF-IDF),[19] Word2Vec[20] and Doc2Vec[21]—and compared their performance with different predictive models (text vectorisation step in figure 1).

TF-IDF measures a word's importance in a text by computing its term frequency, indicating the word's relative frequency in a document.[19] This method is effective for assessing word relevance in document queries. Word2Vec vectorises text using a neural network to create word embeddings, mapping words to vectors.[20] It employs a sliding window technique, using either the continuous bag-of-words (CBOW) method to predict a word from its context or the skip-gram method to predict context words from a given word. Doc2Vec, a generalised Word2Vec, vectorises entire paragraphs or documents directly into single vectors, bypassing the averaging step required in Word2Vec.[21] It offers two algorithms: distributed memory (DM) and distributed bag of words (DBOW).[22] Figure 2 shows the Word2Vec and Doc2Vec algorithms.

Predictive modelling

After the texts were vectorised, the rows of numerically encoded features for both the training and test sets were prepared. We performed feature selection to filter out features that did not positively contribute to the classification task. This step further reduced the feature dimension, resulting in a more compact space. We trained a random forest (RF) classifier to determine the top relevant features for each text vectoriser (feature selection step in figure 1).

The transformed training set was used to train the predictive models using multiple classification methods (classification step in figure 1). We used three different classification approaches: support vector classification (SVC), K-nearest neighbours (KNN) and RF. These approaches spanned a wide variety of classifier categories, including support vector machines, non-parametric methods and ensemble methods, enabling us to evaluate a broader spectrum of model performance. All of these methods were supervised learning techniques; therefore, we used the annotated training set, composed of 70 clinical notes, to train each model.

The SVC finds a hyperplane that maximises the margin between the nearest data points of each label, with hyperparameters tuned for optimal separation.[23] KNN classifies by voting among the 'k' nearest training data points to an input query, leading to larger models with more data.[24 25] RF, an ensemble of decision trees, combines their predictions to reduce overfitting and variance, using moderately tuned hyperparameters for peak performance.[26] We chose the hyperparameter set with moderate parameter tuning to maximise model performance and trained an RF model with the same feature-label pairs from the training set to build a classifier.

We performed a random search combined with fivefold cross-validation to determine the optimal parameters for SVC, KNN and RF methods. Random hyperparameter search randomly selects values from predefined ranges or distributions to evaluate model performance. This is typically done using techniques such as k-fold cross-validation, where the training set is further divided into k-folds, and the model is trained and evaluated on different subsets of data, with each fold used as the validation set once. Then the model is trained and tested multiple times with different hyperparameter values to obtain an estimate of its performance.[27 28]

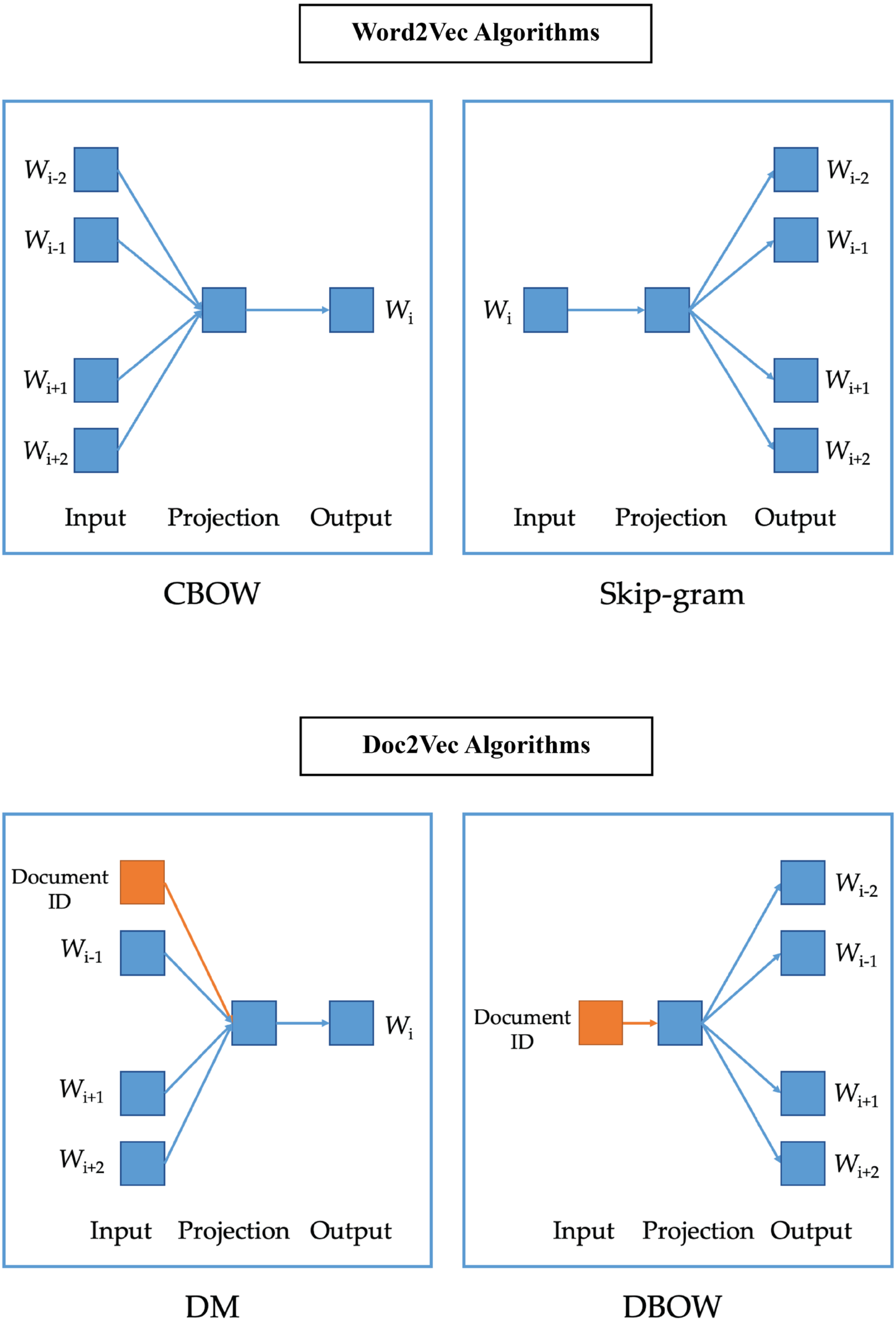

Figure 2 Word2Vec and Doc2Vec algorithms (W_i represents i-th word in a given text). CBOW, continuous bag of words; DBOW, distributed bag of words; DM, distributed memory.

We used the NLTK library[29] for text cleaning, Scikit-Learn for data splitting, TF-IDF vectorisation, predictive modelling (SVC, KNN and RF), random search with cross-validation and evaluation and the Gensim library for Word2Vec and Doc2Vec implementation.[30]

Model evaluation

We used three commonly used performance metrics for model evaluation: precision, recall and area under the receiver operating characteristic curve (AUC). Precision gauges the model's accuracy in predicting positive classes, aiming to reduce false positives. Recall measures the model's success in identifying actual positives, targeting the reduction of false negatives. AUC reflects the model's ability to differentiate between classes across various thresholds, with higher values denoting greater discrimination.

RESULTS

Among the 1000 clinical notes we collected, 100 were randomly selected and annotated by a clinical expert, while the remaining 900 were used to build the text corpus. We found 41 positive notes and 59 negative notes from these 100 annotated notes. We divided the annotated notes into training and test sets (using random selection of 70 and 30 notes, respectively) for modelling. The training set included 27 positive samples, whereas the test set had 14 due to random selection. The annotated dataset comprised 41% of positive labels. The distribution of positive labels was 39% in the training set and 47% in the test set, closely reflecting the entire dataset. We used the 900 unannotated notes and 70 training notes (970 in total) to build our text corpus in the text vectorisation model for the final analysis. We identified 13 029 unique words after the stemming process[18] among the 970 clinical notes selected for training text vectorisation (embedding) model. The mean value was 657.8, and the SD was 438.0. The minimum value recorded was 8, and the maximum was 2721. The 25th percentile was 372.5, the median (50th percentile) was 619.0 and the 75th percentile was 857.8.

We began by using 970 clinical notes as the corpus input for the TF-IDF model, transforming these notes into vectorised features for training and test sets. The n-gram range was set from 1–3 g, resulting in an output feature dimension of 408 791 for the training set. Similarly, we used the same corpus to train a Word2Vec word embedding model, following the TF-IDF approach. After training, each word in a note was converted into a vector, and each note was represented by the average of these vectors.

For the Doc2Vec approach, we trained a word-embedding model with the same set of clinical notes, treating each note as a document in the Doc2Vec framework. This enabled us to infer document vectors for each note, which were then used in training predictive models. Both Word2Vec and Doc2Vec models were assigned a feature size of 2000. In the Word2Vec model, the CBOW approach was preferred over Skip-gram due to its superior performance, while for the Doc2Vec model, we chose the DM model over the DBOW method. A window size of three was selected for both models. The hyperparameters for these models are detailed in table 1.

Feature selection is a crucial step in machine learning model development, as it helps identify the most relevant features or variables that contribute to a model's prediction performance. We employed a selection-by-model approach for feature selection after training the vectorisers. In this method, an intermediate model is trained to rank the importance of features based on their impact on the overall accuracy or performance of the model. Specifically, we trained an intermediate RF classifier to rank the importance of features based on their contribution to maximising the accuracy of the classifier. The RF classifier was chosen for its ability to handle non-linearity, interactions and most importantly, its ability to provide feature importance estimation. Then we selected the top 300 features ranked by the RF classifier across all text vectorisation models to balance between capturing relevant information and avoiding overfitting or issues with high-dimensional data.

We performed a random search with fivefold cross-validation to determine the optimal parameters for each model. The hyperparameters used in the random search are listed in table 1. The random search keeps the

Table 1 Text vectoriser classifiers hyperparameters for each text vectorisation model

Text vectoriser hyperparameters		
TF-IDF	n-gram range: 1–3; max document frequency: 1.0; min document frequency count: 1	
Word2Vec	Features size: 2000; window size: 3; min count: 1; training algorithm: CBOW; training epochs: 20	
Doc2Vec	Features size: 2000; window size: 3; min count: 1; training algorithm: distributed memory; training epochs: 20	
Classifier hyperparameters		
SVC	Kernel: type: RBF, inverse regularisation coefficient: 1.0	
KNN	TF-IDF	Number of neighbours: 3, leaf size: 10
	Word2Vec	Number of neighbours: 10, leaf size: 10
	Doc2Vec	Number of neighbours: 3, leaf size: 10
RF	TF-IDF	Number of estimators: 50, max tree depth: 10
	Word2Vec	Number of estimators: 50, max tree depth: 5
	Doc2Vec	Number of estimators: 50, max tree depth: 5

KNN, K-nearest neighbours; RBF, radial basis function; RF, random forest; SVC, support vector classification; TF-IDF, term frequency-inverse document frequency.

best-performing model from the fivefold validation, and we used the cached model for subsequent evaluations.

We used a test set comprising 30 annotated clinical notes to evaluate the models. These clinical notes were annotated with ground truth labels, serving as the reference for evaluating the model's predictions. We calculated precision, recall, F1-score, accuracy and AUC for each trained model using the test set and used these performance metrics to assess the model's performance. These metrics provide quantitative measures of the model's performance and can aid in selecting the best performing model for the given classification task. The results can be found in table 2, where combination of text vectorisation and classification models were evaluated using specific metrics along with their 95% CI. We measured the CI using the bootstrapping method, with 1000 iterations of sampling.

The TF-IDF results indicated the highest AUC performance when combined with SVC (0.82), followed by RF (0.82) and KNN (0.73) on the test set. However, the Word2Vec model failed to train effectively with SVC, as indicated by zero scores in both precision and recall. For KNN (0.58) and RF (0.57), the AUC was also low compared with other vectorisation methods. In contrast, the Doc2Vec results showed the highest AUC when paired with RF (0.90), followed by SVC (0.86) and KNN (0.57). Notably, the Doc2Vec-RF combination achieved the best AUC results across all combinations. The performance of Word2Vec was lower than that of other text vectorisers, and KNN was generally less effective than other classifiers, except when used with Word2Vec. Although we used k-fold cross-validation for hyperparameter tuning, the RF results from the training set suggested overfitting. Interestingly, the Doc2Vec-RF combination showed a narrower gap between training and test set results across all metrics. Figure 3 illustrates the ROC curves for text vectorization and classification methods (figure 3).

DISCUSSION

The main goal of this study was to develop an end-to-end NLP pipeline for extracting treatment outcomes of breast cancer among women from under-represented populations, aiming to obtain important insights to enhance care for these populations. By focusing on these groups, our study sought to fill a critical knowledge gap and contribute to fostering equity in healthcare treatment outcomes.

We designed and implemented a systematic and automated approach that leverages NLP techniques to extract relevant information from clinical notes and accurately classify the extracted texts. We compared several algorithms to assess the efficiency of each approach. Specifically, this project holds significant value because it employed algorithms to analyse the treatment outcomes of patients with breast cancer from under-represented populations. These groups have been previously understudied, leading to a gap in our understanding of how treatments affect them differently. By employing NLP to analyse clinical notes, we gained a more comprehensive understanding of the optimal algorithms for extracting treatment outcomes for patients with breast cancer from under-represented populations. This approach has the potential to lead to more equitable healthcare outcomes in these communities.

The development of this NLP system involved consideration of two key components: text vectorisation and classification. We compared and evaluated different text vectorisation methods (TF-IDF, Word2Vec and Doc2Vec) in combination with classification models (SVC, KNN and RF). The results indicated that both the TF-IDF and Doc2Vec text vectorisation models demonstrated the highest performance in terms of AUC when combined with the RF classification model. This suggests that these two vectorisation methods were effective in capturing the relevant information from the clinical notes data and improving the performance of the classification model. In comparison, the SVC and KNN classification models performed worse in terms of AUC when combined with the TF-IDF and Doc2Vec vectorisation methods. The fact that the TF-IDF and Doc2Vec models outperformed the Word2Vec model in our specific task suggests that performing vectorisation at the document level, as opposed to individual words, is crucial for building a stable and accurate clinical note classifier. The simple mean vector approach, where individual feature vectors from the words in a document are averaged to obtain a document-level representation, used in Word2Vec, was not suitable for the clinical notes in an EHR system.

Among the different classification algorithms we evaluated, the RF classifier demonstrated the best performance in most of the comparisons. This suggests that the underlying structure of the 300-feature space used in our study was non-linear, and the reduction of variation achieved through ensemble learning in RF contributed to better model training. This finding aligns well with our expectations, considering the complexity of clinical notes data and the relatively large size of the feature vector used in our study.

However, we also observed that the recall scores of the RF model were relatively lower compared with precision, indicating that the model had more false negatives. In other words, it tended to miss some positive cases, leading to lower recall rates. The same trend is also observable in other methods, indicating this is not a classifier-specific problem. Instead, this could be due to the imbalanced nature of the dataset, or the specific characteristics of the clinical notes being analysed. Further investigation is needed to understand the reasons behind this observation and identify potential ways to improve the recall performance of the classification model. On the other hand, KNN model showed the lowest performance in terms of recall compared with the SVC and RF models. This could be attributed to the fact that KNN is an instance-based model, which is more susceptible to noise in the data. Perhaps the clinical notes in our study data might have

Table 2 Performance comparison of text vectorisation and classification methods on both training and test datasets (each metric is shown with its 95% CI)

	Training set														
	SVC					KNN					RF				
	P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC
TF-IDF	0.96 (0.88 to 1.00)	1.00 (1.00 to 1.00)	0.98 (0.94 to 1.00)	0.99 (0.96 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	0.37 (0.18 to 0.56)	0.54 (0.32 to 0.71)	0.76 (0.66 to 0.86)	0.95 (0.91 to 0.98)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)
Word2Vec	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.61 (0.50 to 0.71)	0.65 (0.49 to 0.80)	0.83 (0.60 to 1.00)	0.37 (0.19 to 0.55)	0.51 (0.29 to 0.70)	0.73 (0.63 to 0.83)	0.77 (0.66 to 0.87)	0.95 (0.82 to 1.00)	0.67 (0.48 to 0.83)	0.78 (0.63 to 0.90)	0.86 (0.77 to 0.93)	0.96 (0.91 to 0.99)
Doc2Vec	1.00 (1.00 to 1.00)	0.85 (0.71 to 0.96)	0.92 (0.83 to 0.98)	0.94 (0.89 to 0.99)	1.00 (0.99 to 1.00)	1.00 (1.00 to 1.00)	0.41 (0.23 to 0.60)	0.58 (0.36 to 0.75)	0.77 (0.67 to 0.87)	0.97 (0.93 to 0.99)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)
	Test set														
	SVC					KNN					RF				
	P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC	P	R	F1	Acc.	AUC
TF-IDF	0.62 (0.38 to 0.87)	0.71 (0.46 to 0.93)	0.67 (0.44 to 0.84)	0.67 (0.47 to 0.80)	0.82 (0.62 to 0.97)	1.00 (1.00 to 1.00)	0.29 (0.07 to 0.55)	0.44 (0.13 to 0.70)	0.67 (0.50 to 0.83)	0.73 (0.56 to 0.89)	0.80 (0.50 to 1.00)	0.57 (0.30 to 0.83)	0.67 (0.42 to 0.86)	0.73 (0.57 to 0.87)	0.80 (0.64 to 0.94)
Word2Vec	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.53 (0.37 to 0.70)	0.77 (0.56 to 0.93)	0.57 (0.14 to 1.00)	0.29 (0.07 to 0.54)	0.38 (0.10 to 0.62)	0.57 (0.40 to 0.730	0.58 (0.37 to 0.77)	0.62 (0.25 to 1.00)	0.36 (0.13 to 0.62)	0.45 (0.13 to 0.69)	0.60 (0.43 to 0.77)	0.57 (0.32 to 0.79)
Doc2Vec	1.00 (1.00 to 1.00)	0.50 (0.25 to 0.80)	0.67 (0.40 to 0.87)	0.77 (0.60 to 0.90)	0.86 (0.72 to 0.96)	0.33 (0.00 to 1.00)	0.07 (0.00 to 0.25)	0.12 (0.00 to 0.35)	0.50 (0.33 to 0.67)	0.57 (0.40 to 0.77)	0.89 (0.62 to 1.00)	0.57 (0.31 to 0.82)	0.70 (0.40 to 0.88)	0.77 (0.60 to 0.90)	0.90 (0.76 to 0.99)

Acc, accuracy; AUC, area under the curve; F1, F1-score; KNN, K-nearest neighbours; P, precision; R, recall; RF, random forest; SVC, support vector classification; TF-IDF, term frequency-inverse document frequency.

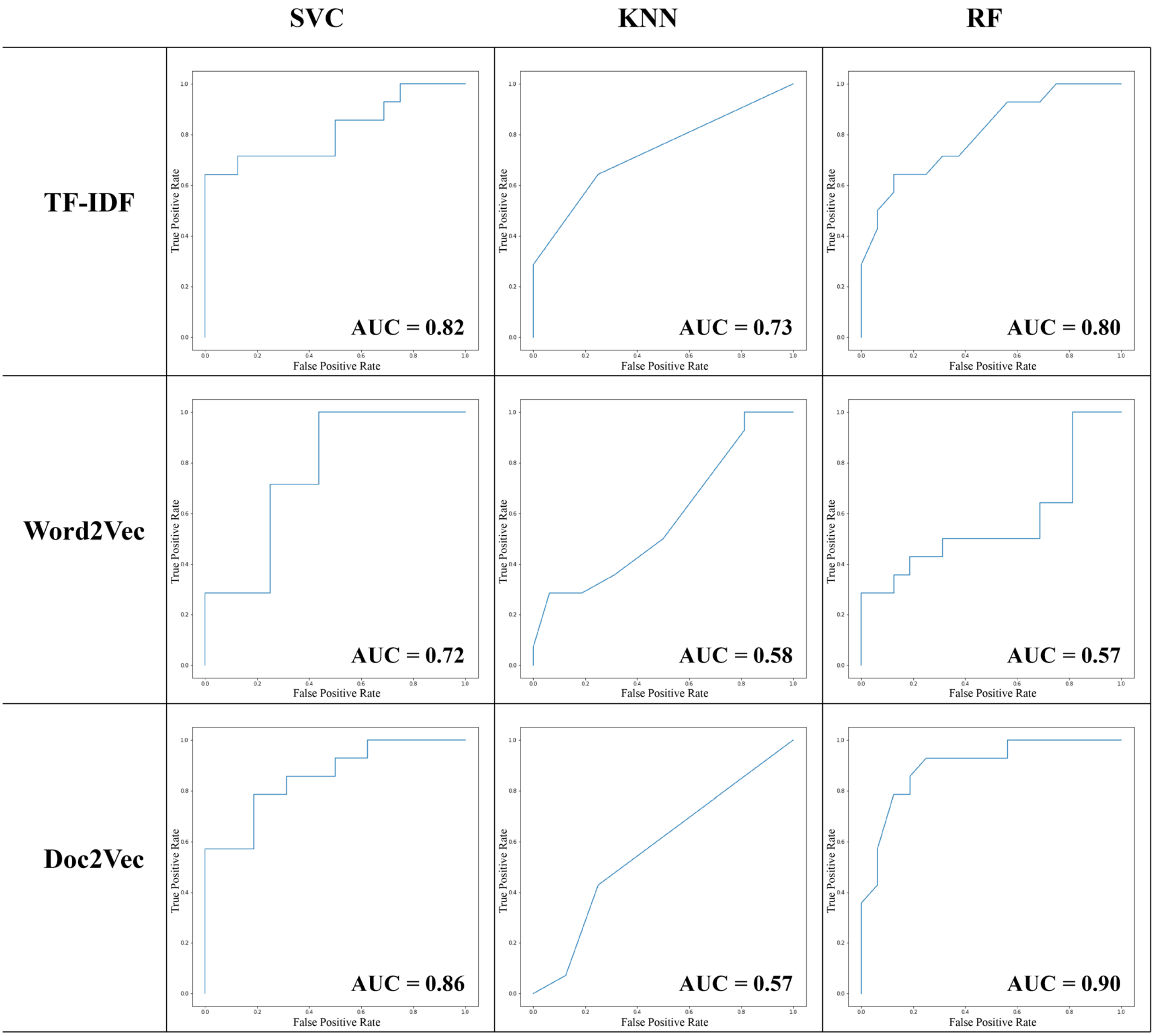

Figure 3 Receiver operating characteristic curves for text vectorisation and classification methods. AUC, area under the curve; KNN, K-nearest neighbours; RF, random forest; SVC, support vector classification; TF-IDF, term frequency-inverse document frequency.

contained noise or outliers that affected the performance of the KNN model negatively.

Our study has some limitations. Due to the small size of annotated notes, our approach has limited generalisability. We aim to collect more data and annotations to address this in the future work. Expanding our dataset will allow us to better validate our findings, refine our methodology and potentially increase the accuracy and robustness of our predictions. Additionally, we observed that document-level approaches, especially the deep learning-based Doc2Vec model, performed better than other methods in general. We plan to explore the possibility of using embeddings from large language models for text vectorisation to further improve performance.

Overall, our study contributes to the field of clinical NLP by developing a high-performing pipeline for capturing invasive breast cancer treatment outcomes of women from under-represented populations. While the NLP methods we employed were not new in themselves, their application to our specific target demographic sets our work apart. We were able to access and interpret a wealth of nuanced, unstructured data that would otherwise have been difficult to investigate. We could identify potential disparities in care, offering valuable insights that can be used to develop strategies for achieving more equitable healthcare outcomes for these vulnerable groups. In addition, our findings provided insights into the importance of document-based text vectorisation and the efficacy of ensemble methods in the context of clinical notes data. Furthermore, the pipeline we developed is adaptable and generalisable to other NLP tasks that involve different clinical note classifications, based on its fully automated end-to-end design. This suggests that the approach we developed has potential for broader applications in various clinical NLP tasks beyond breast cancer treatment outcomes.

CONCLUSION

In this study, we developed a high-performance NLP pipeline that accurately discerns treatment outcomes of invasive breast cancer in under-represented women, highlighting previously overlooked disparities in care. Emphasising the significance of document-based text vectorisation, our method notably leveraged the TF-IDF and Doc2Vec models. Coupled with the superior performance of ensemble methods, especially the RF classifier, we could effectively navigate complex clinical notes. Despite challenges like lower recall rates in some classifiers, the adaptable design of our pipeline signifies its potential for broader clinical NLP applications beyond just breast cancer outcomes. Future research should validate its scalability and generalisability across diverse healthcare datasets.

Contributors JIP is responsible for the overall content as guarantor. The conceptualisation of the study was led by JIP and DK. The methodology was developed by JIP and DK. The validation of the study's findings was conducted by JIP and JWP. The formal analysis and investigation were carried out by JIP. JIP and KZ provided the essential resources. Data curation was handled by JIP. The original draft was prepared by JIP, with review and editing contributions from JWP, DK and KZ. Visualisation efforts were undertaken by JIP and DK. Supervision and project administration were overseen by JIP. All authors have read and agreed to the published version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the Institutional Review Board at University of California, Irvine.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

Jung In Park http://orcid.org/0000-0002-1771-7361

REFERENCES

1 Siegel RL, Miller KD, Fuchs HE, *et al*. Cancer statistics. *CA A Cancer J Clinicians* 2021;71:7–33.

2 Yedjou CG, Sims JN, Miele L, *et al*. Health and racial disparity in breast cancer. *Adv Exp Med Biol* 2019;1152:31–49.

3 Bickell NA, Wang JJ, Oluwole S, *et al*. Missed opportunities: racial disparities in adjuvant breast cancer treatment. *J Clin Oncol* 2006;24:1357–62.

4 American Cancer Society. Breast cancer facts & figures 2019-2020. 2020. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

5 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58.

6 Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol* 2020;145:463–9.

7 Koleck TA, Dreisbach C, Bourne PE, *et al*. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019;26:364–79.

8 Dreisbach C, Koleck TA, Bourne PE, *et al*. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* 2019;125:37–46.

9 Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J Biomed Inform* 2019;100:103301.

10 Topaz M, Murga L, Gaddis KM, *et al*. Mining fall-related information in clinical notes: comparison of rule-based and novel word embedding-based machine learning approaches. *J Biomed Inform* 2019;90:103103.

11 Fernandes MB, Valizadeh N, Alabsi HS, *et al*. Classification of neurologic outcomes from medical notes using natural language processing. *Expert Syst Appl* 2023;214:119171.

12 Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med Inform Decis Mak* 2019;19:71.

13 Weng W-H, Wagholikar KB, McCray AT, *et al*. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak* 2017;17:155.

14 Banerjee I, Bozkurt S, Caswell-Jin JL, *et al*. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO Clin Cancer Inform* 2019;3:1–12.

15 Carrell DS, Halgrim S, Tran D-T, *et al*. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol* 2014;179:749–58.

16 Wang H, Li Y, Khan SA, *et al*. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artif Intell Med* 2020;110:101977.

17 Bird S. NLTK: the natural language Toolkit. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions; 2006:69–72.

18 Runowicz CD, Leach CR, Henry NL, *et al*. American Cancer Society/American society of clinical oncology breast cancer survivorship care guideline. *CA Cancer J Clin* 2016;66:43–73.

19 Ramos J. Using TF-Idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning; 2003:29–48.

20 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. *arXiv* [Preprint] 2013.

21 Le Q, Mikolov T. Distributed representations of sentences and documents. International conference on machine learning; 2014:1188–96.

22 Bilgin M, Senturk IF. Sentiment analysis on Twitter data with semi-supervised Doc2Vec. 2017 International Conference on Computer Science and Engineering (UBMK); Ieee, 661–6. Antalya.

23 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273–97.

24 Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory* 1967;13:21–7.

25 Aha DW, Kibler D, Albert MK. Instance-based learning Algorithms. *Mach Learn* 1991;6:37–66.

26 Ho TK. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition; 1995:278–82.

27 Hardeniya N, Perkins J, Chopra D, *et al*. Natural language processing: python and NLTK. Packt Publishing Ltd; 2016.

28 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.

29 Řehůřek R, Sojka P. Gensim—statistical semantics in python. 2011. Available: genism.org

30 Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 2018;19:270.