Classification of kinetic-related injury in hospital triage data using NLP

—

Supplementary Material

Midhun Shyam, Jim Basilakis, Kieran Luken, Steven Thomas, John Crozier, Paul M Middleton, X. Rosalind Wang

1 Introduction

This is the Supplementary Material for the paper "Classification of kinetic-related injury in hospital triage data using NLP", as published in the Proceedings of The 21st International Conference on Advanced Data Mining and Applications, 2025 (ADMA 2025).

This Supplementary Material is organised as follows: Firstly, provides additional information on the data and preprocessing (Section 2). Secondly, detailed results of fine-tuning the model (Step 1 as described in the main paper, Section 3), prediction using the fine-tuned models (Step 2, Section 4), domain adaptation (Step 3, Section 5), and prediction using the domain-adapted models (Step 4, Section 6). Finally, we provide results on the computing resources used with different number of CPUs (Section 7).

The three fine-tuning strategies for the Neural Network (NN) used in this work are:

1. **Neural Network 1 (NN1)** — The weights from the Bio-Clinical BERT (BCB) were used *as is* (frozen) and the fine-tuning was used to optimise the Classification Head (CH) parameters only.

2. **Neural Network 2 (NN2)** — All but the last layer of the BCB were frozen, and the fine-tuning process was used to optimise the weights of the CH and the final encoder layer (layer 12).

3. **Neural Network 3 (NN3)** — Fine-tuning to optimise the weights for layers 11 and 12 of the BCB and the CH.

2 Data and Preprocessing

We used data from two sources: Medical Information Mart for Intensive Care (MIMIC)-III and The Comprehensive Epidemiological Database for Research, Innovation and Collaboration (CEDRIC):

2.1 MIMIC-III data

The MIMIC-III Clinical Database [3], publicly available through PhysioNet[1] [2] provides free access to extensive physiological and clinical data alongside open-source software. The data comprises observations with presenting problems of over 40,000 critical care patients at the Beth Israel Deaconess Medical Center, Boston, Massachusetts Emergency Department (ED), between 2001 and 2012. We used the data collected from the critical care unit, specifically the table `NOTEEVENTS` that contains various de-identified clinical notes, for this paper. These data contains over two million samples of clinical notes, medical reports and discharge summaries, and is the same data set that the BCB was originally trained on [1].

For this work, we focused only on the free text data contained within the `TEXT` column. The `TEXT` data contains a variety of notes, such as "History of Present Illness", "Chief Complaint", "Admit Diagnosis" etc. For this work, only "History of Present Illness" notes were used, as it is most similar to triage notes. These notes were cleaned and truncated to 512 tokens, with encodings, special characters, empty lines, duplicates, recurring sentences within the same observation, and other irrelevant elements removed. This process reduced the dataset down to 50,546 observations (henceforth called "cleaned `NOTEEVENTS`"). Examples of critical care units clinical notes, similar to what is in this final set, are shown on the top table of Table 1.

We created a dataset — henceforth called "*MIMIC data*" — using the following steps: First, we used Regular Expression (REGEX) to identify any notes in the cleaned `NOTEEVENTS` data related to positive cases that contain keywords specific to kinetic energy-related vehicular injuries (as listed under "Vehicle" in Table 2). This step classified 2034 observations as positive kinetic injury cases and 48,512 as other. Second, the subset of 2034 true cases underwent manual review, which identified 725 false cases. These are notes which had one or more of the keywords in the text, but were not patient presentations due to kinetic-related vehicular injuries.

Third, to create a balanced dataset of equal numbers of positive and negative cases, we then added 583 notes to this data set, from the rest of the cleaned `NOTEEVENTS` data. These cases were randomly selected, but care was taken to ensure they followed the same word length distribution as the true cases. These samples also underwent manual review, where 176 were identified as possible true cases and were thus excluded. Therefore, we created a dataset of 1309 positive kinetic-related vehicular injury cases and 1132 negative cases.

2.2 CEDRIC data

The Comprehensive Epidemiological Database for Research, Innovation and Collaboration (CEDRIC) database is held at South Western Emergency Research Institute (SWERI) within the Ingham Institute of Applied Medical Research [2] it captures and links data from ED presentations in South West Sydney Local Health District (SWSLHD)[3] since 2005.

For this work, we concentrated on the ED triage comments in CEDRIC (See bottom table of Table 1 for example triage notes). We performed a REGEX word search of the database using the expressions listed in Table 2 for positive and negative samples. To

[1] a research resource founded in 1999 under the National Institutes of Health (NIH)

[2] https://sweri.com.au/

[3] Comprised of: Bankstown-Lidcombe, Bowral, Camden and Campbelltown, Fairfield, and Liverpool Hospitals.

Table (1) Example of synthetic ICU (top) and triage (bottom) clinical notes for kinetic-related vehicular trauma cases (+ve label) and otherwise (-ve label). The corresponding rows in the tables show the exact same case, as would have been written by the two different departments.

Label	Example ICU clinical notes (similar to MIMIC data)
+ve	X, a 36-year-old marketing executive, was brought to the emergency department at 7:45 PM following a collision with a power pole while riding her electric bike. Witnesses reported she was travelling at approximately 30 km/h when she swerved to avoid a pedestrian and struck the pole chest-first. Upon arrival, she was conscious and oriented but visibly distressed, with vital signs showing tachycardia (HR 118), tachypnoea (RR 24), normal blood pressure (BP 112/74), and oxygen saturation of 97% on room air. She complained of severe central chest pain radiating to her back, rating it 8/10, which worsened with inspiration. Primary survey revealed intact airway and breathing with no obvious external chest injuries, though auscultation noted muffled heart sounds. She was immediately transferred to the resuscitation bay where an eFAST scan revealed a significant pericardial effusion suggestive of cardiac tamponade. Urgent CT confirmed a moderate-sized pericardial effusion (2.3cm) with associated findings of cardiac compression. No aortic injury was identified, but a small right haemothorax was noted. Upon return to the resuscitation bay, ...
-ve	An 80-year-old man, Mr. X, presented to the emergency department at 7:30 PM with confusion, fever, and severe lower abdominal pain. His daughter reported that he had been increasingly lethargic over the past 48 hours, with decreased oral intake and had become confused this evening. She noted he had complained of burning during urination for approximately three days but had refused to see his doctor. On examination, Mr. X was disoriented (Glasgow Coma Scale 13/15), febrile (temperature $39.2°C$), tachycardic (heart rate 118 bpm), hypotensive (BP 82/45 mmHg), and tachypnoeic (respiratory rate 28/min). He appeared pale and diaphoretic with delayed capillary refill of 4 seconds. His oxygen saturation was 94% on room air. ...

Label	Example triage notes (similar to CEDRIC Data)
+ve	36yo female presents post electric bike vs. power pole collision approx. 30 minutes ago. Alert but anxious, reports severe central chest pain 8/10, worse on inspiration. Denies LOC at scene. No helmet worn. O/E RR 24, SpO2 97% RA, HR 118, BP 112/74, Temp 36.7°C. Skin clammy. No visible chest wall deformity but reporting pain on palpation of sternum. Multiple abrasions to right arm and leg. Known allergies to penicillin (rash). Not pregnant. No regular medications.
-ve	88yo man brought in by family, c/o abdominal pain, fever and suprapubic pain. Daughter also states that her father has been complaining for burning when he passes urine. PMHx of prostatism, hypertension, and T2DM. O/E Pale and sweaty, RR 28, SpO2 94% RA, temp 39.2°C, HR 118 bpm, BP 82/45, GCS 13/15. Confused and disorientated in time, place and person.

Table (2) Terms used in REGEX pattern matching search. The terms identified for 'vehicle' are those where possible kinetic injury took place. These terms are used to find the positive data in both databases. The terms identified for 'trauma' are general trauma cases that are not specific to kinetic impact; these terms are only used to find non-kinetic related cases in the CEDRIC database.

Label	**Terms used in `regex` text matching**
Vehicle	mva, mba, vehicle, bus, pedestrian, passenger, ute, ped, bike, dirtbike, motorbike, pushbike, scooter, truck, bicycle, motorcycle, driver, driving, rtc, rta, `\d*km[a-zA-Z/]*`, skateboard, surfing, surf, horse, collision, crossing, buggy, ebike, jetski, vs car, car vs, car accident, moving car, traffic light, traffic lights, hit by car, hit by a car, car hit, airbag, airbags, T boned, reversed, struck by car, struck by a car, tyre
Trauma	`(?<!nil )`trauma, `(?<!nil )`injury, `(?<!nil )`injured, `(?<!nil )`injuries, strain, rupture, sprain, fracture, dislocation, damage, contusion, tear, assault, accident, homicide, blast, split, bruised, laceration, concussion, lac, crush, broken, fall, fallen, #

avoid duplicates in search results, we searched for samples in the different cases from two different years in the database. Furthermore, as we know that REGEX will return a small percentage of false positive samples, and the number of negative samples greatly outnumbers positive samples in the database, we used terms from both cases to search for notes in the second year. To ensure we obtained a balanced data set, we returned the same number of samples from the two years in the database. Both data sets were manually inspected and labelled.

We thus created two data sets:

1. CEDRIC One: positive cases from 2022 and both cases from 2021, resulting in 447 kinetic injury cases and 553 others.

2. CEDRIC Two: positive cases from 2023 and both cases from 2022, resulting in 413 kinetic injury cases and 587 others.

3 Fine-tuning on MIMIC data

The results after fine-tuning the Bio-Clinical BERT Classifier (BCBC) configurations using the NN1, NN2, and NN3 fine-tuning strategies for the AdamW optimiser are presented in Figure 1, and for the Adam and SGD optimisers in Figures 2 and 3 respectively. Each figure contains three sub-figures, with (a) showing the accuracy, (b) the F1-Score, and (c) the total training time in minutes. The results are also presented in tabular form for NN1 (Table 3), NN2 (Table 4), and NN3 (Table 5).

Figures 1–3 demonstrate the different error metrics tested across each NN architecture. We found that most learning/dropout rate combinations provided similar results, with the exception of the NN3 configuration using $lr = 0.005$, which had a significantly worse accuracy and F1-score. Excluding the models optimised using SGD, and the largest learning rate ($lr = 0.005$), the NN1 architecture performed worst across all metrics, with NN2 and NN3 performing similarly. We found that NN3 architecture, using the Adam optimiser with $lr = 0.0001$ and $dr = 0.15$ provided the best error metrics (accuracy of $95.0 \pm 0.9\%$, and F1-score of $95.4 \pm 0.8\%$). However, this was at the expense of training time, as the AdamW optimiser was consistently quicker, with training times generally taking between a third and half as long as the Adam optimiser (Table 5, bottom), for marginally worse accuracy and F1-score.

We performed a two-sampled t-test ($N = 10$ for the number of repetitions of each model configurations trained) of all pairs of results for accuracy and F1-score in Tables 3-5 ten times. We observed that about half of the more than 300 pairs had statistically significant differences between the results at $p = 0.05$. The distinction between significant and non-significant difference was subtle — e.g. 0.945 ± 0.008 vs. 0.933 ± 0.010 was not significant, while 0.947 ± 0.013 vs. 0.933 ± 0.010 was significant at $p = 0.05$. However, in a practical sense, there was no real difference between accuracies of $95.0 \pm 0.9\%$ (the best Adam configuration) and $94.2 \pm 1.2\%$ (the worst AdamW configuration, shown in Table 5(top)) even though they were statistically significantly different.

Table (3) Accuracy, F1-score and training time on GPU of fine-tuning on MIMIC data with the different optimiser using NN1, and at different learning and drop out rates.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.902 ± 0.011	0.903 ± 0.012	0.903 ± 0.012	0.903 ± 0.012	0.903 ± 0.011	0.903 ± 0.012
AdamW	0.902 ± 0.013	0.903 ± 0.013	0.904 ± 0.012	0.908 ± 0.013	0.908 ± 0.011	0.906 ± 0.010
SGD	0.808 ± 0.025	0.816 ± 0.023	0.802 ± 0.028	0.872 ± 0.013	0.872 ± 0.013	0.872 ± 0.010
	F1-score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.910 ± 0.010	0.911 ± 0.011	0.911 ± 0.010	0.912 ± 0.011	0.911 ± 0.009	0.911 ± 0.011
AdamW	0.911 ± 0.011	0.912 ± 0.011	0.912 ± 0.011	0.915 ± 0.011	0.916 ± 0.010	0.914 ± 0.009
SGD	0.838 ± 0.020	0.845 ± 0.018	0.835 ± 0.022	0.885 ± 0.012	0.884 ± 0.011	0.885 ± 0.009
	Time (Min) (Learning rate, Drop out rate)					
Adam	52.012 ± 31.871	54.622 ± 36.095	53.690 ± 33.143	24.953 ± 11.361	28.247 ± 13.267	29.620 ± 12.057
AdamW	60.175 ± 36.935	57.892 ± 34.549	58.152 ± 35.111	28.470 ± 17.422	30.002 ± 21.738	26.483 ± 17.515
SGD	104.448 ± 43.670	104.433 ± 43.664	104.443 ± 43.626	94.610 ± 40.999	104.432 ± 43.617	103.918 ± 43.134

Table (4) Accuracy, F1-score and training time on GPU of fine-tuning on MIMIC data with the different optimiser using NN2, and at different learning and drop out rates.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.943 ± 0.011	0.944 ± 0.010	0.942 ± 0.014	0.943 ± 0.015	0.945 ± 0.014	0.942 ± 0.012
AdamW	0.942 ± 0.007	0.945 ± 0.009	0.945 ± 0.012	0.945 ± 0.010	0.944 ± 0.012	0.944 ± 0.006
SGD	0.852 ± 0.021	0.850 ± 0.012	0.846 ± 0.019	0.925 ± 0.012	0.920 ± 0.013	0.922 ± 0.013
	F1-score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.947 ± 0.010	0.948 ± 0.010	0.946 ± 0.013	0.947 ± 0.014	0.949 ± 0.013	0.946 ± 0.011
AdamW	0.946 ± 0.007	0.949 ± 0.008	0.949 ± 0.011	0.949 ± 0.009	0.948 ± 0.012	0.948 ± 0.006
SGD	0.867 ± 0.019	0.865 ± 0.012	0.861 ± 0.017	0.932 ± 0.011	0.926 ± 0.012	0.929 ± 0.011
	Time (Min) (Learning rate, Drop out rate)					
Adam	10.450 ± 8.321	8.648 ± 2.085	10.335 ± 4.967	7.272 ± 2.570	9.092 ± 3.768	11.217 ± 3.999
AdamW	5.588 ± 0.549	6.293 ± 2.313	6.208 ± 2.604	6.108 ± 2.612	4.797 ± 0.670	5.293 ± 1.869
SGD	106.710 ± 44.473	115.078 ± 43.474	106.702 ± 44.418	96.905 ± 39.984	85.037 ± 39.787	88.497 ± 44.997

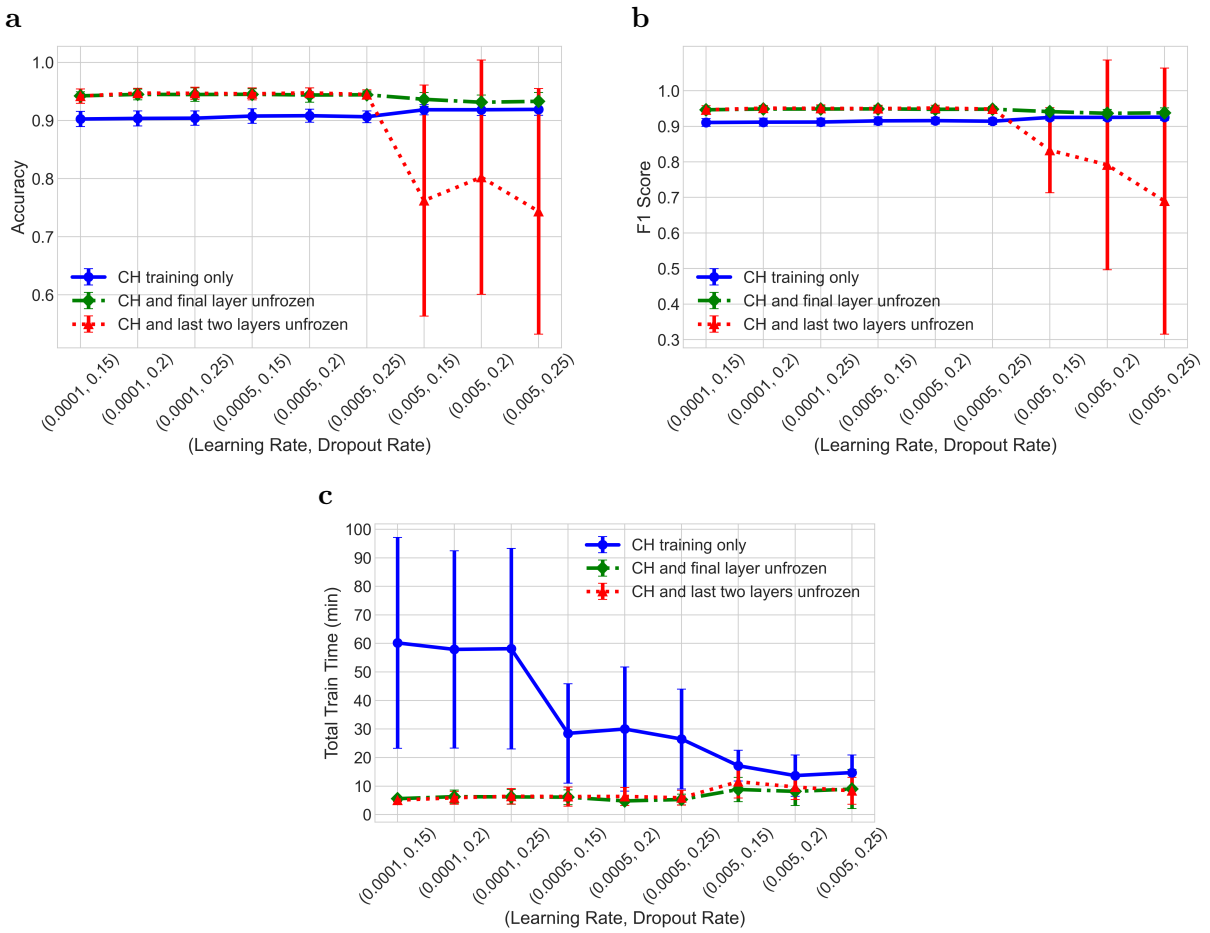

Figure (1) Results of fine-tuning using the AdamW optimiser: (a) Accuracy of the validation data; (b) F-1 score on the validation data; (c) training time.

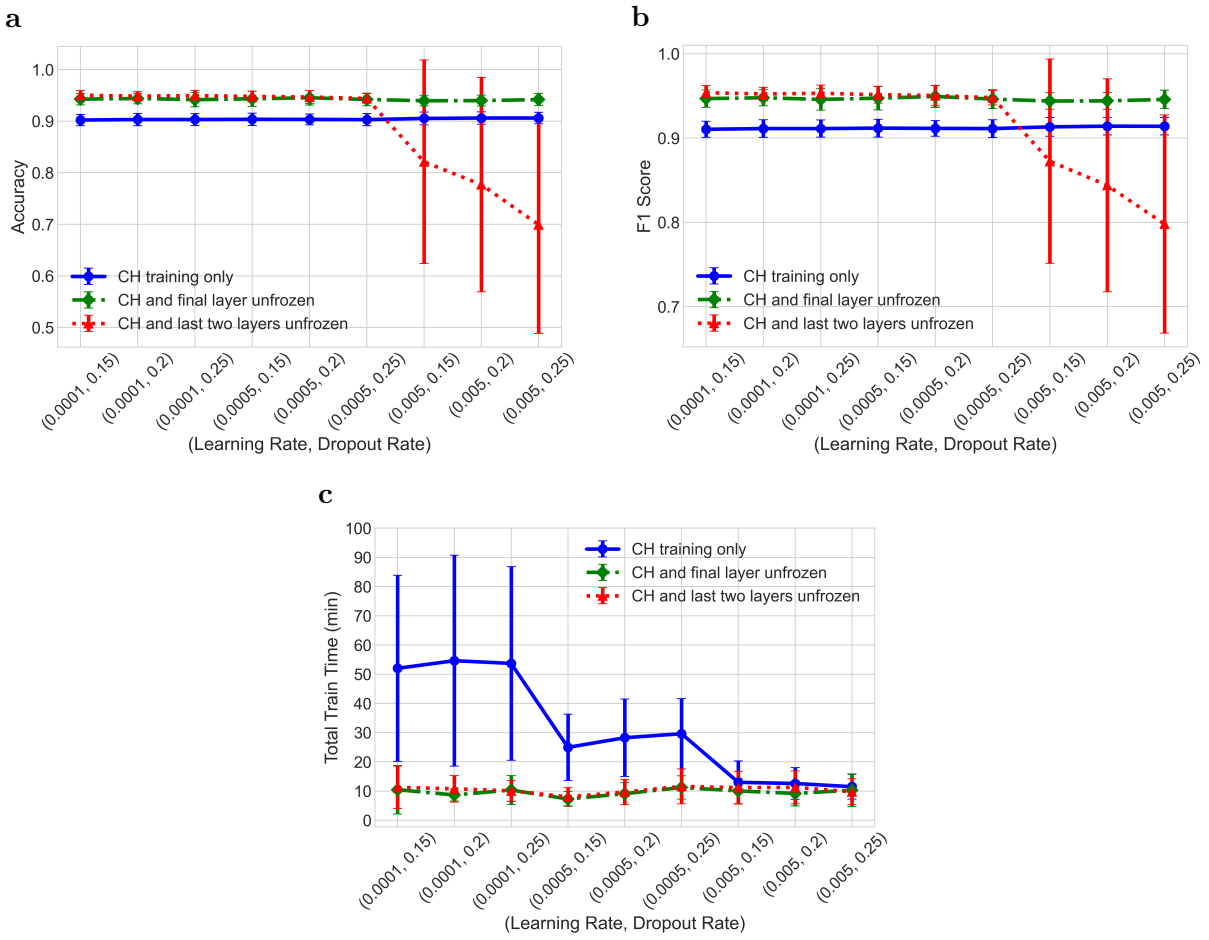

Figure (2) Results of fine-tuning using the Adam optimiser: (a) Accuracy of the validation data; (b) F-1 score on the validation data; (c) training time.

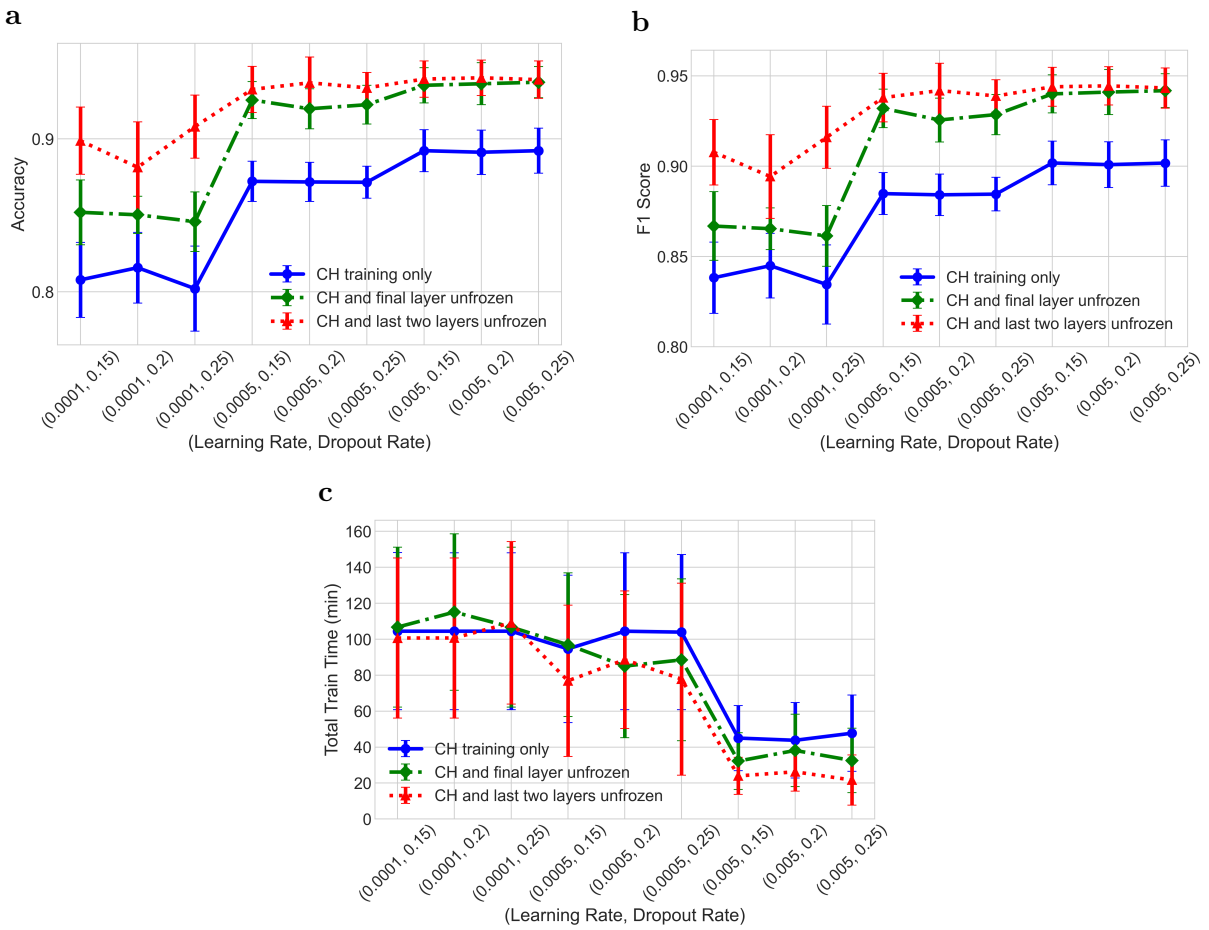

Figure (3) Results of fine-tuning using the SGD optimiser: (a) Accuracy of the validation data; (b) F-1 score on the validation data; (c) training time.

Table (5) Results of fine-tuning NN3 using the different optimisers and at different hyperparameters. The best result for each optimiser is highlighted.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	**0.950 ± 0.009**	0.948 ± 0.009	0.949 ± 0.010	0.948 ± 0.011	0.947 ± 0.013	0.944 ± 0.009
AdamW	0.942 ± 0.012	0.947 ± 0.008	0.947 ± 0.010	0.946 ± 0.010	**0.947 ± 0.009**	0.945 ± 0.008
SGD	0.899 ± 0.022	0.881 ± 0.030	0.908 ± 0.021	0.932 ± 0.015	**0.937 ± 0.017**	0.933 ± 0.010
	F1-score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	**0.954 ± 0.008**	0.952 ± 0.008	0.953 ± 0.010	0.951 ± 0.010	0.950 ± 0.012	0.948 ± 0.009
AdamW	0.946 ± 0.011	0.951 ± 0.008	0.950 ± 0.010	0.950 ± 0.009	**0.951 ± 0.008**	0.949 ± 0.008
SGD	0.908 ± 0.018	0.894 ± 0.023	0.916 ± 0.017	0.938 ± 0.013	**0.942 ± 0.015**	0.939 ± 0.009
	Time (min) (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	11.267 ± 7.288	10.760 ± 4.515	10.070 ± 3.622	**8.022 ± 3.161**	9.663 ± 4.318	11.598 ± 5.960
AdamW	**5.057 ± 0.505**	5.905 ± 2.276	6.430 ± 2.689	6.327 ± 3.350	6.372 ± 3.175	5.898 ± 2.603
SGD	100.635 ± 44.524	100.660 ± 44.502	109.112 ± 45.218	**76.850 ± 42.045**	88.600 ± 38.255	77.797 ± 53.375

4 Prediction with fine-tuned models

The prediction results from the fine-tuned models are presented in Figures 4-5 and Tables 6-8. We found that the results mirror those in Section 3, with the Adam and AdamW optimised models providing roughly the same accuracies and F1-scores. Like Section 3, the SGD optimised models, and $lr = 0.005$ perform the worst.

We found the model that provides the best accuracy and F1-score is the AdamW optimised NN3 architecture, using $lr = 0.0005$ and $dr = 0.15$ — a similar configuration to Section 3 — giving an accuracy of $83.9 \pm 1.8\%$, and an F1-score of $80 \pm 2.5\%$. Similar to Section 3, about half of the accuracy and F1-score pairs reported exhibited a statistically significantly difference between the results at $p = 0.05$. However, as noted in Section 3, the practical difference for clinicians between a model producing an accuracy of 83.9% (AdamW optimised NN3 with $lr = 0.0005$ and $dr = 0.15$) and another with an accuracy of 82.0% (AdamW optimised NN3 with $lr = 0.0001, dr = 0.15$) is minimal.

We note that the drop in accuracy from $\sim 95\%$ reported Section 3 to $\sim 84\%$ is not unexpected, as the fine-tuned model has been fine-tuned with data from one context (the MIMIC data). This section then took those trained models, and tested them against the CEDRIC Two data — a fundamentally different context, as seen in Table 1.

The total prediction time as shown in subfigure (c) of Figure 4-6 was similar across all configurations, which is not surprising. All fine-tuned models took between 170 and 290 seconds to predict the data set of 1000 triage notes. Hence, this task could easily be achieved on much larger datasets within medical facilities. SWSLHD receives around half a million patients in ED in a year, this means even with minimum computing resources the classification process will take just one day for a year's data.

Given the significant under-performance of the NN1 architecture, the SGD optimised models, and the models with a learning rate of 0.005, we did not continue further testing with these models.

Table (6) Prediction results on CEDRIC Two using fine-tuned NN1 with the different optimisers and at different learning and drop out rates.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.672 ± 0.008	0.672 ± 0.009	0.673 ± 0.007	0.683 ± 0.009	0.682 ± 0.004	0.686 ± 0.009
AdamW	0.675 ± 0.015	0.682 ± 0.011	0.682 ± 0.010	0.695 ± 0.012	0.696 ± 0.010	0.693 ± 0.008
SGD	0.521 ± 0.034	0.516 ± 0.024	0.519 ± 0.024	0.598 ± 0.012	0.596 ± 0.012	0.599 ± 0.016
	F1-score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.612 ± 0.013	0.612 ± 0.015	0.611 ± 0.015	0.628 ± 0.010	0.629 ± 0.008	0.630 ± 0.011
AdamW	0.616 ± 0.019	0.622 ± 0.012	0.621 ± 0.012	0.641 ± 0.010	0.640 ± 0.009	0.636 ± 0.009
SGD	0.561 ± 0.032	0.566 ± 0.018	0.558 ± 0.045	0.553 ± 0.015	0.543 ± 0.020	0.558 ± 0.019

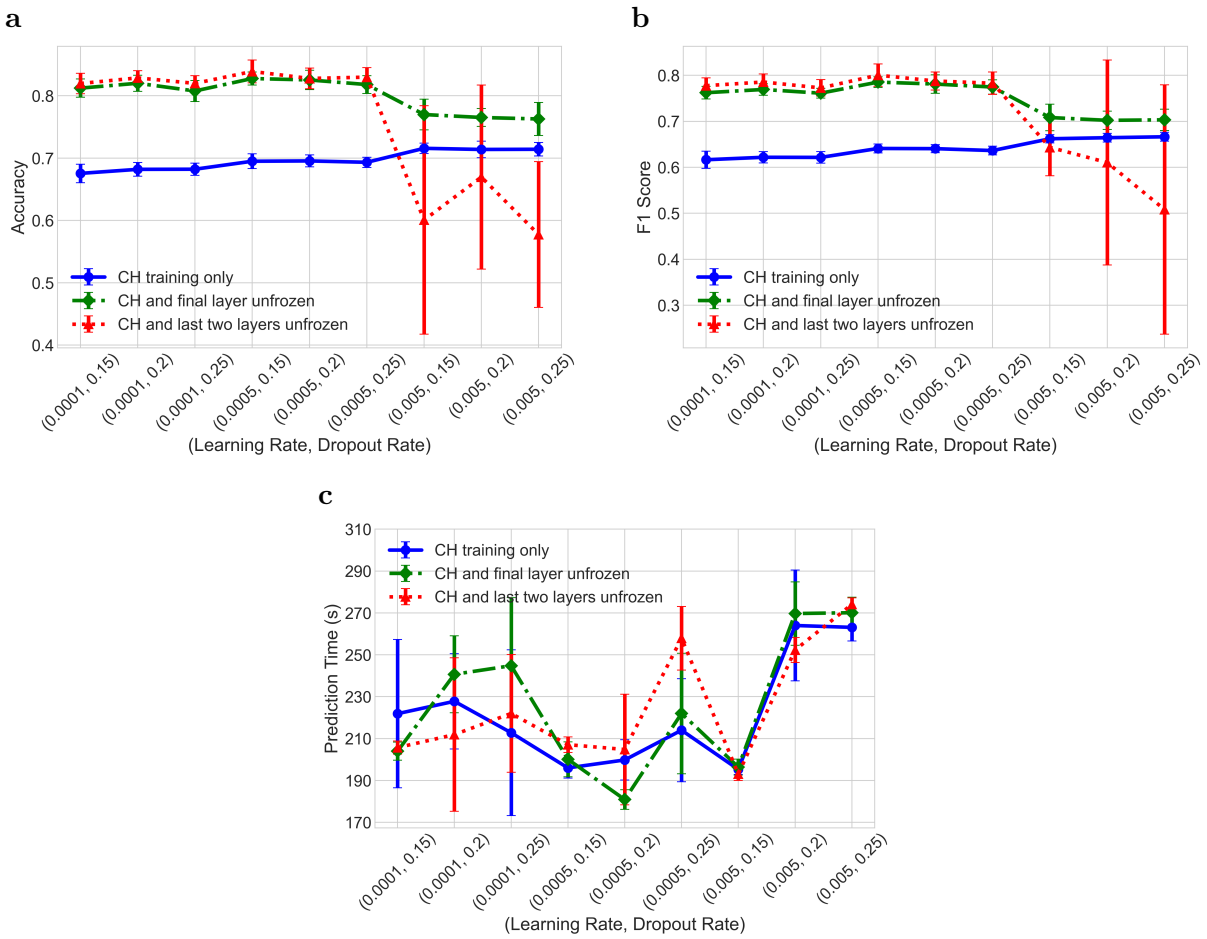

Figure (4) Results of prediction on CEDRIC Two using models fine-tuned with the AdamW optimiser: (a) Prediction accuracy; (b) F-1 score; (c) Prediction on CPU time (s).

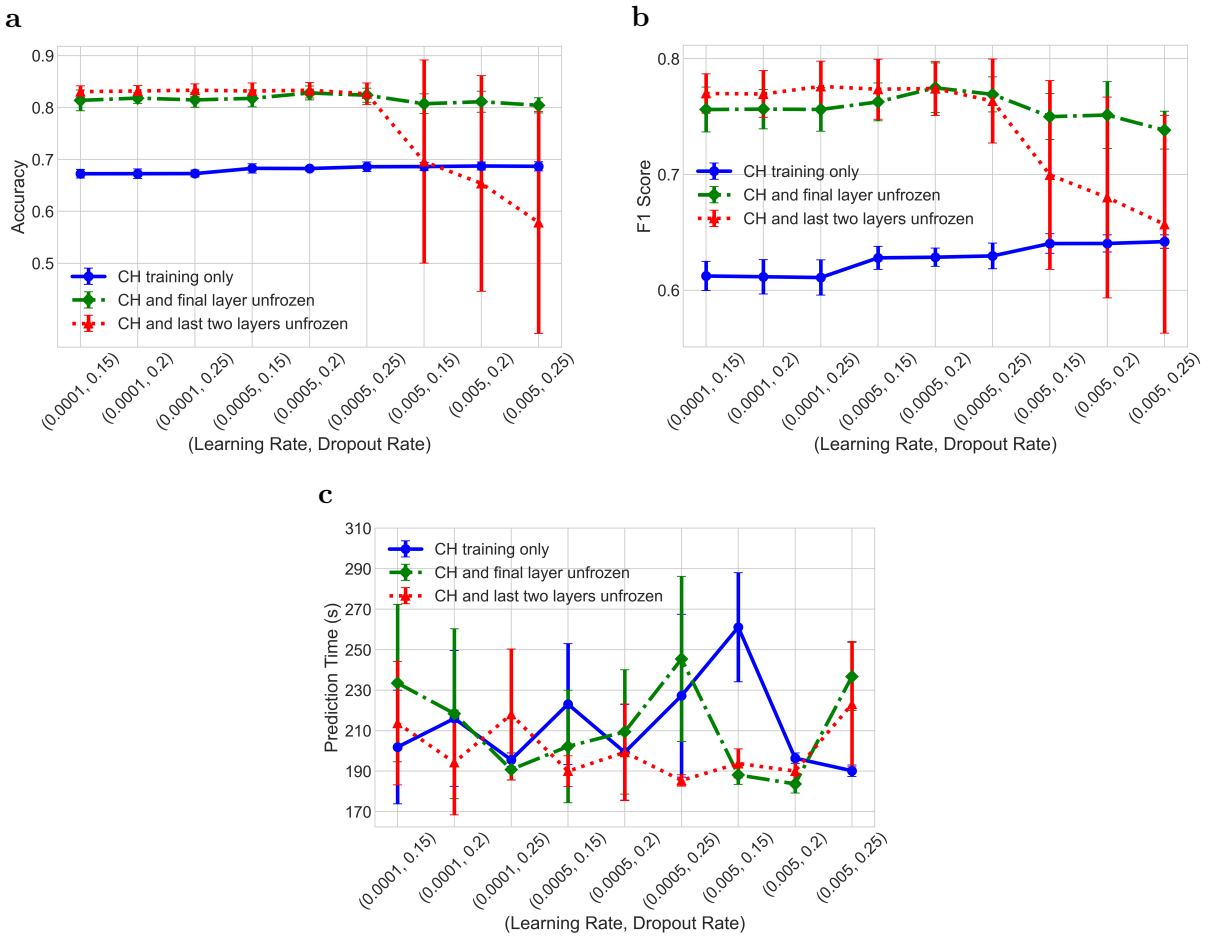

Figure (5) Results of prediction on In-house Two using models fine-tuned with the Adam optimiser: (a) Prediction accuracy; (b) F-1 score; (c) Prediction on CPU time (s).

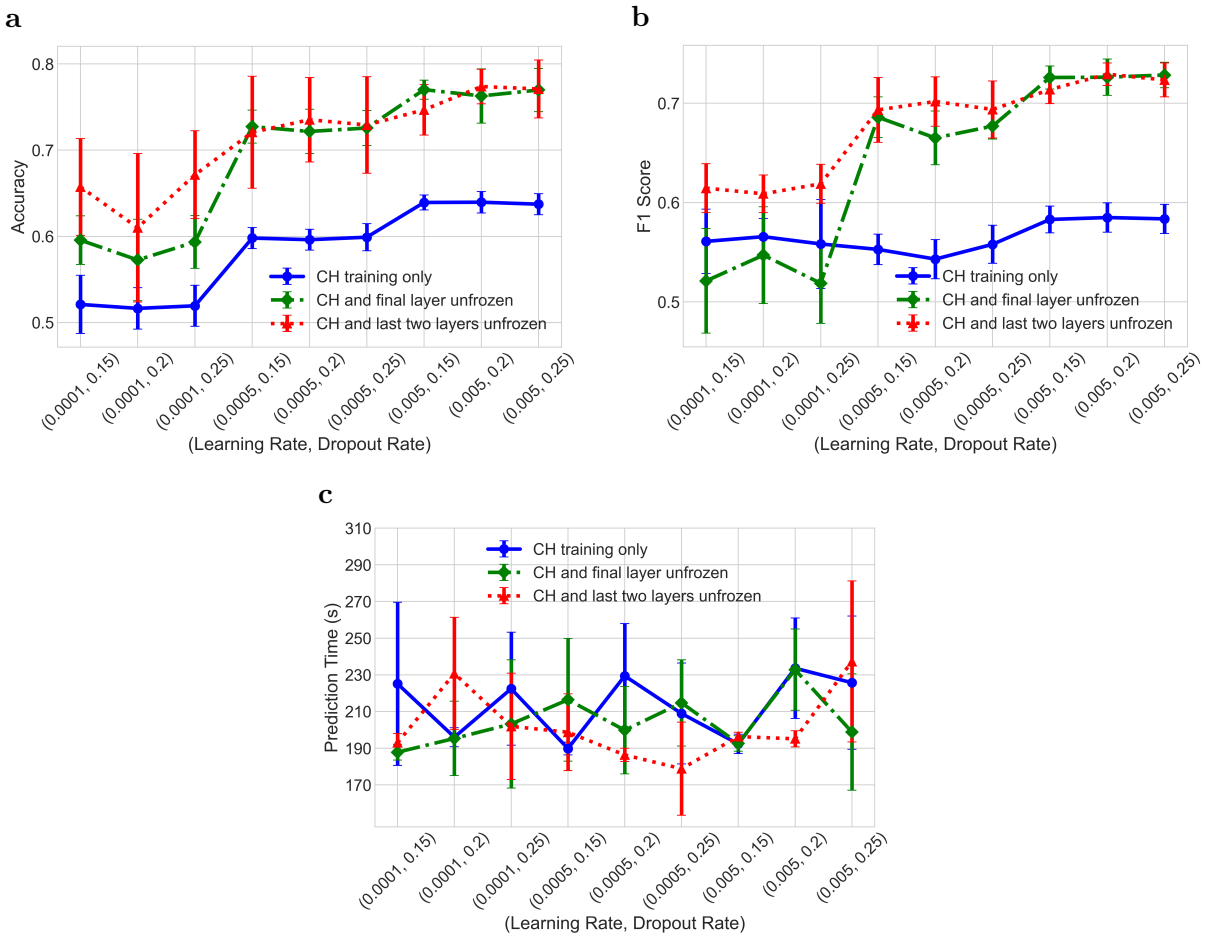

Figure (6) Results of prediction on In-house Two using models fine-tuned with the SGD optimiser: (a) Prediction accuracy; (b) F-1 score; (c) Prediction on CPU time (s).

Table (7) Prediction results on CEDRIC Two using fine-tuned NN2 using the different optimisers and at different learning and drop out rates.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.814 ± 0.020	0.818 ± 0.010	0.815 ± 0.014	0.818 ± 0.016	0.828 ± 0.013	0.824 ± 0.014
AdamW	0.812 ± 0.015	0.820 ± 0.013	0.808 ± 0.017	0.828 ± 0.010	0.825 ± 0.016	0.818 ± 0.014
SGD	0.596 ± 0.028	0.573 ± 0.047	0.593 ± 0.031	0.727 ± 0.019	0.722 ± 0.026	0.726 ± 0.020
	F1-score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.756 ± 0.019	0.756 ± 0.017	0.756 ± 0.019	0.763 ± 0.016	0.775 ± 0.021	0.769 ± 0.015
AdamW	0.762 ± 0.014	0.769 ± 0.013	0.761 ± 0.011	0.785 ± 0.012	0.781 ± 0.020	0.774 ± 0.016
SGD	0.521 ± 0.053	0.547 ± 0.049	0.519 ± 0.040	0.686 ± 0.020	0.665 ± 0.027	0.677 ± 0.013

Table (8) Accuracy and F1-score of prediction on CEDRIC Two data using the fine-tuned NN3 models. The best result for each optimiser is highlighted.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.831 ± 0.011	0.832 ± 0.010	$\mathbf{0.833 \pm 0.012}$	0.832 ± 0.015	0.833 ± 0.015	0.826 ± 0.021
AdamW	0.820 ± 0.016	0.829 ± 0.012	0.820 ± 0.012	$\mathbf{0.839 \pm 0.018}$	0.828 ± 0.017	0.830 ± 0.015
SGD	0.657 ± 0.056	0.610 ± 0.086	0.671 ± 0.051	0.721 ± 0.065	$\mathbf{0.735 \pm 0.049}$	0.729 ± 0.056
	F1 Score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.770 ± 0.017	0.769 ± 0.020	$\mathbf{0.776 \pm 0.022}$	0.774 ± 0.026	0.774 ± 0.023	0.763 ± 0.036
AdamW	0.777 ± 0.017	0.785 ± 0.018	0.773 ± 0.018	$\mathbf{0.800 \pm 0.025}$	0.788 ± 0.020	0.783 ± 0.024
SGD	0.615 ± 0.025	0.609 ± 0.019	0.619 ± 0.020	0.693 ± 0.033	$\mathbf{0.702 \pm 0.025}$	0.694 ± 0.029

5 Further fine-tuning on CEDRIC One Data

The tabulated results of the models fine-tuned (domain adapted) on CEDRIC One data are shown in Tables 9 and 10. The results are also plotted in Figures 7 and 8 for models fine-tuned with the Adam and AdamW optimisers respectively.

Similar to Sections 3 and 4, we found the Adam optimised models provided the best overall performance with the best accuracy (using $lr = 0.0005, dr = 0.2$, gave an accuracy of $94.5 \pm 0.9\%$) and F1-score (using $lr = 0.0001, dr = 0.15$, and gave an F1-score of $94.0 \pm 1.2\%$). We note that this configuration provided the equal best average accuracy, however had a larger standard deviation than the other model identified.

Unlike the previous sections, all pairs of results reported were *not* statistically significantly different from each other (at $p = 0.05$). Once again, however, the practical difference between an accuracy of 94.5% (the best performing model overall) and 93.5% (the worst performing model) is not significant.

The fine-tuning time on the CPU was higher, as expected. Given the exact same model configuration and data, fine-tuning would have taken only a small percentage of the time on the GPU facility. However, since the total training time was less than three hours on a CPU with limited RAM, the model was well within the realm of feasibility for adaptation in any medical facilities.

Table (9) Accuracy, F1-score and training time on CPU of further fine-tuning (domain adaptation) on CEDRIC One data with the different optimiser using NN2, and at different learning and drop out rates.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.936 ± 0.008	0.929 ± 0.012	0.928 ± 0.011	0.938 ± 0.009	0.943 ± 0.011	0.940 ± 0.009
AdamW	0.923 ± 0.011	0.932 ± 0.015	0.922 ± 0.013	0.934 ± 0.013	0.941 ± 0.013	0.934 ± 0.006

	F1 Score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.930 ± 0.009	0.923 ± 0.013	0.922 ± 0.011	0.932 ± 0.009	0.937 ± 0.013	0.934 ± 0.009
AdamW	0.916 ± 0.011	0.927 ± 0.015	0.916 ± 0.014	0.928 ± 0.013	0.936 ± 0.014	0.928 ± 0.007

	Time (min) (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	198.092 ± 59.439	184.190 ± 48.655	185.488 ± 45.923	131.315 ± 40.023	127.185 ± 63.776	98.440 ± 21.057
AdamW	111.487 ± 32.506	104.337 ± 24.244	178.830 ± 14.122	166.618 ± 32.676	148.200 ± 20.756	93.328 ± 21.506

Table (10) Accuracy, F1-score and training time on CPU of further fine-tuning on CEDRIC One data with the AdamW optimiser using NN3. The best result for each optimiser is highlighted.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.945 ± 0.011	0.942 ± 0.012	0.940 ± 0.009	0.940 ± 0.013	$\mathbf{0.945 \pm 0.009}$	0.943 ± 0.012
AdamW	0.936 ± 0.009	$\mathbf{0.945 \pm 0.017}$	0.935 ± 0.011	0.938 ± 0.011	0.937 ± 0.012	0.936 ± 0.011

	F1 Score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	$\mathbf{0.940 \pm 0.012}$	0.936 ± 0.013	0.934 ± 0.010	0.933 ± 0.014	0.939 ± 0.011	0.936 ± 0.013
AdamW	0.930 ± 0.010	$\mathbf{0.940 \pm 0.018}$	0.929 ± 0.012	0.931 ± 0.012	0.931 ± 0.013	0.930 ± 0.012

	Time (min) (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	207.720 ± 45.481	156.333 ± 47.129	156.617 ± 46.909	165.522 ± 29.733	$\mathbf{96.982 \pm 10.188}$	111.652 ± 29.873
AdamW	$\mathbf{96.703 \pm 17.070}$	163.218 ± 27.705	160.700 ± 26.273	159.258 ± 16.115	101.602 ± 34.954	114.013 ± 48.953

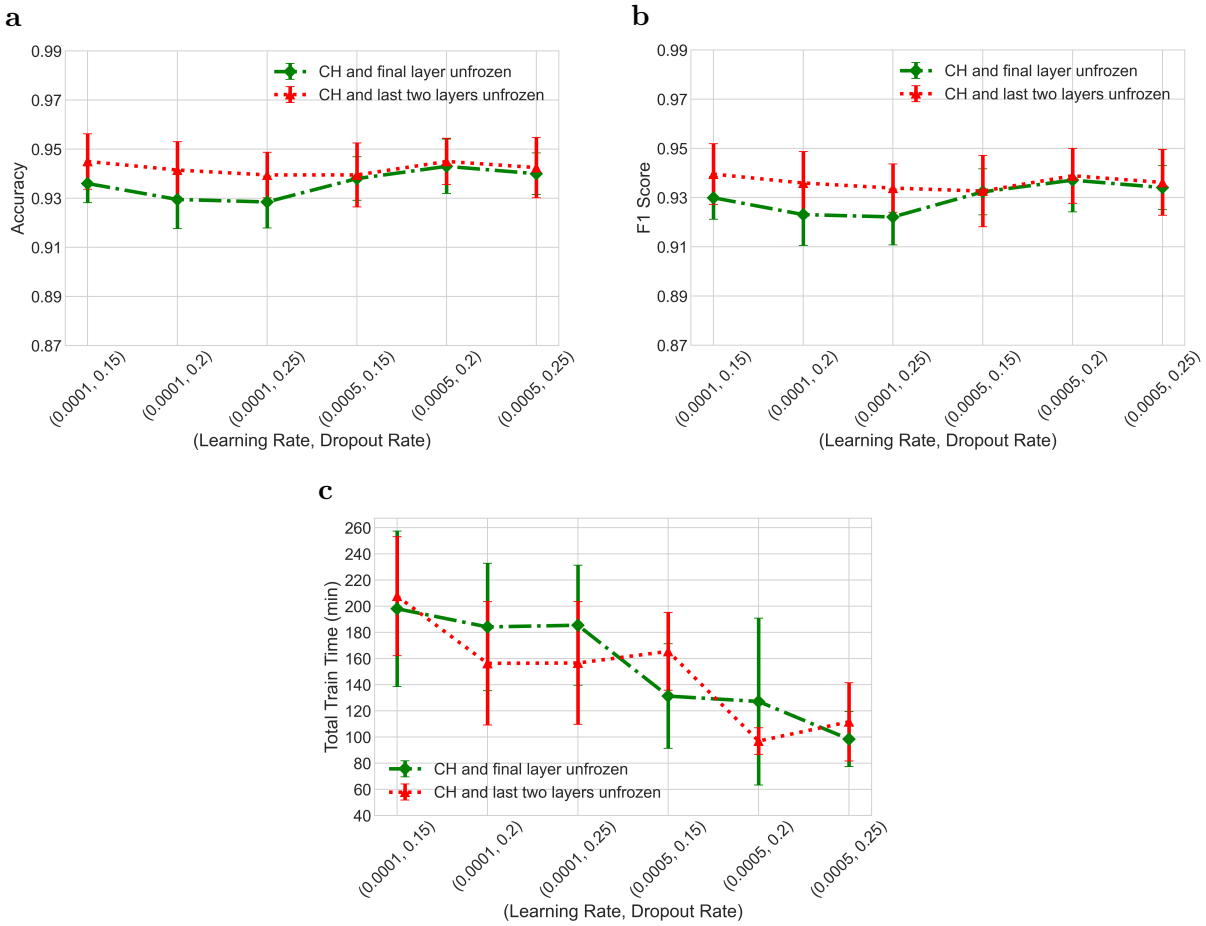

Figure (7) Results of further fine-tuning on CEDRIC One data with the Adam optimiser: (a) Prediction accuracy; (b) F-1 score; (c) Training time on CPU (min).

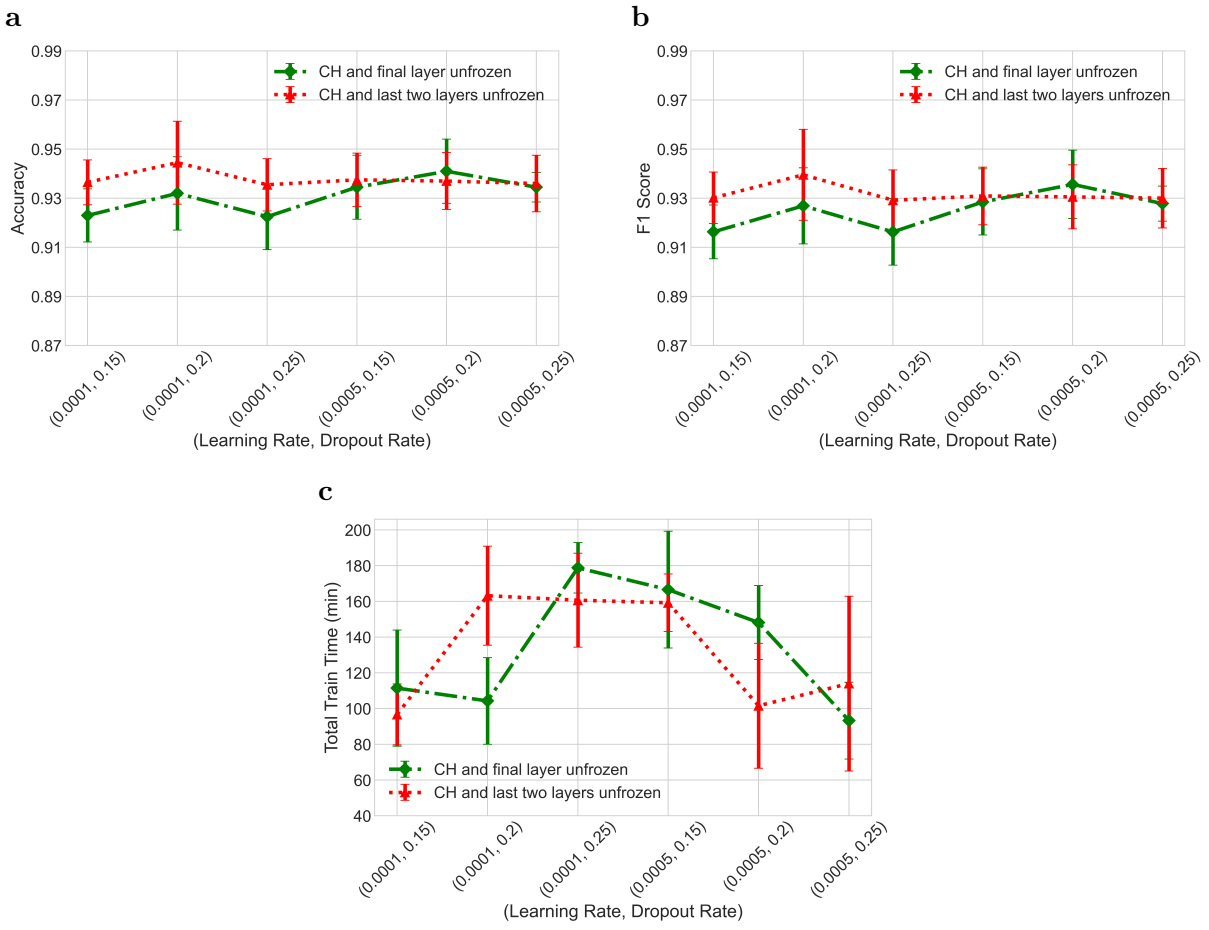

Figure (8) Results of further fine-tuning (domain adaptation) on CEDRIC One data with the AdamW optimiser: (a) Prediction accuracy; (b) F-1 score; (c) Training time on CPU (min).

6 Prediction with domain adapted models

Results for prediction on CEDRIC Two data using the NN2 model further fine-tuned (domain adapted) on CEDRIC One data are shown in Table 11. The plots of models fine-tuned with the Adam optimiser are shown in Figure 9. The other results are presented in the main paper.

Table (11) Accuracy, F1-score and training time on CPU of further fine-tuning (domain adaptation) on CEDRIC One data with the different optimiser using NN2, and at different learning and drop out rates.

	Accuracy (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.933 ± 0.004	0.931 ± 0.005	0.930 ± 0.004	0.933 ± 0.005	0.935 ± 0.004	0.932 ± 0.006
AdamW	0.925 ± 0.006	0.923 ± 0.005	0.923 ± 0.007	0.930 ± 0.003	0.931 ± 0.005	0.928 ± 0.005

	F1 Score (Learning rate, Drop out rate)					
Optimiser	(0.0001, 0.15)	(0.0001, 0.2)	(0.0001, 0.25)	(0.0005, 0.15)	(0.0005, 0.2)	(0.0005, 0.25)
Adam	0.922 ± 0.005	0.920 ± 0.006	0.919 ± 0.005	0.922 ± 0.005	0.924 ± 0.004	0.920 ± 0.006
AdamW	0.913 ± 0.007	0.911 ± 0.005	0.911 ± 0.007	0.918 ± 0.004	0.919 ± 0.006	0.916 ± 0.006

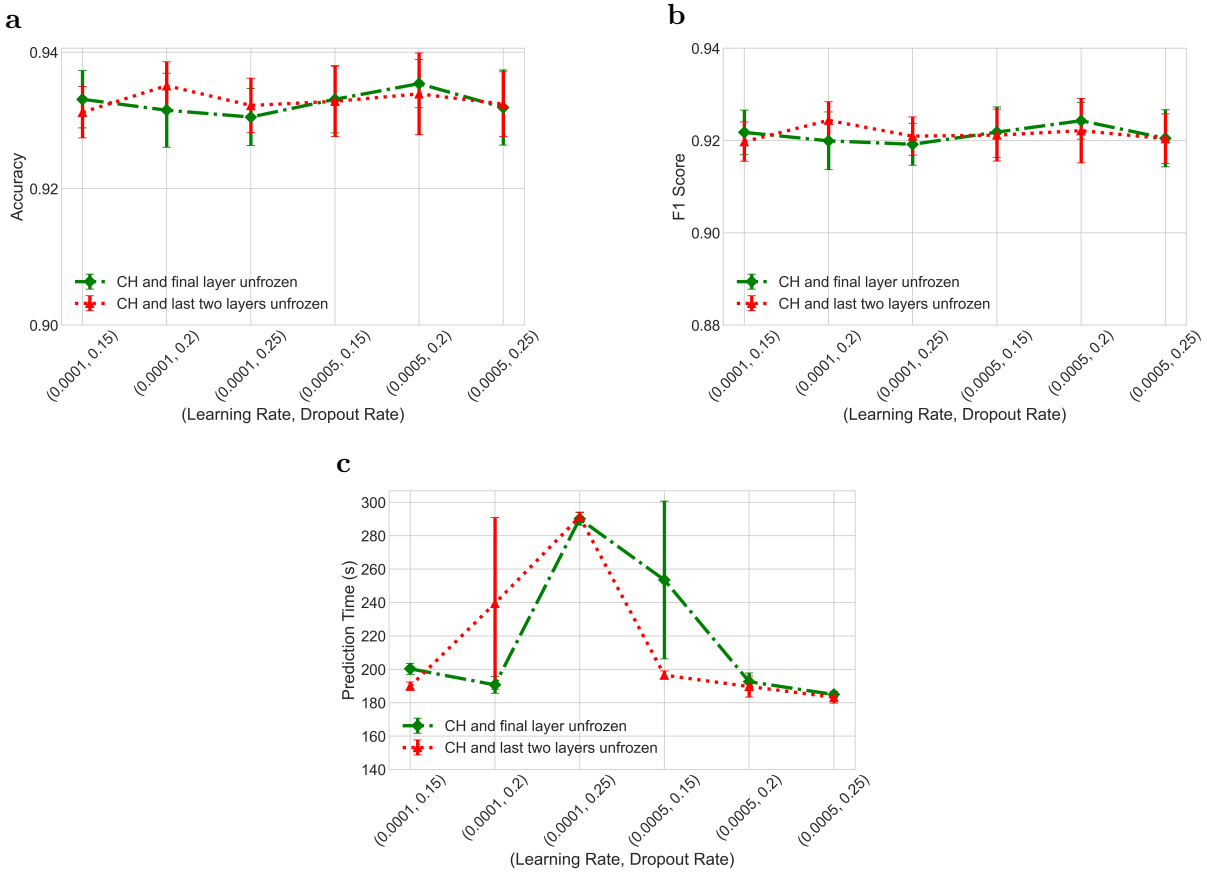

Figure (9) Results of prediction on CEDRIC Two data with the domain adapted models optimised with Adam optimiser: (a) Prediction accuracy; (b) F-1 score; (c) Time taken on CPU (s)

7 Run time on different number of CPUs

We ran the further fine-tuning step using AdamW optimiser with 0.0001 learning rate and 0.15 drop out rate on CEDRIC One data with different number of CPUs.

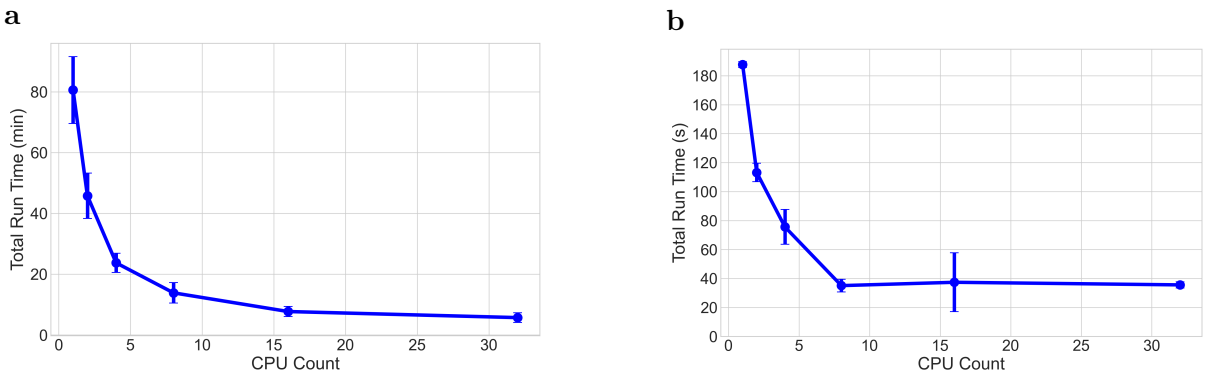

Figure (10) Results of testing the run time with different number of CPUs: (a) further fine-tuning time (min) using CEDRIC One data; (b) prediction time (s) using CEDRIC Two data.

Note that for the prediction time for 16 CPUs, one run took 3 times as long as the other nine runs, hence the large error bar. Also note that for the prediction time for large number of CPUs (with the total run time under a minute), a large percentage of the time would have been spent on data I/O, importing the many packages and libraries, and initialising the model. Hence, the run time is no longer decreasing with the number of CPUs. One would assume that the run time would be lower for large number of CPUs if the size of the data is much larger.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical NLP Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. ACL.

[2] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[3] Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III clinical database (v1.4), 2016.