PENGINTEGRASIAN DATA SEJARAH DAN ANALISIS DATA MASA NYATA UNTUK RAMALAN DAN PEMANTAUAN WABAK PENYAKIT

Tow Ding Feng, Shahnorbanun Sahran

Fakulti Teknologi & Sains Maklumat

Universiti Kebangsaan Malaysia

43600 Bangi, Selangor

Abstrak

Projek ini memfokuskan pada ramalan wabak COVID-19 di peringkat negeri di Malaysia dengan memanfaatkan teknik pembelajaran mesin. Aspek utama projek termasuk penyepaduan data epidemiologi pelbagai sumber, pemodelan siri masa dan kebolehtafsiran ramalan model. Masalah utama yang ditangani ialah kesukaran untuk menghasilkan ramalan kes COVID-19 yang tepat, tepat pada masanya dan setempat, terutamanya apabila wabak itu beralih kepada fasa endemik. Cabaran timbul daripada ketidakkonsistenan data merentas sumber, kesan dasar yang berkembang, dan keperluan untuk pemodelan dinamik yang mampu mengendalikan kebergantungan temporal yang kompleks. Projek ini mencadangkan pembangunan rangka kerja ramalan yang teguh menggunakan model *Temporal Fusion Transformers*,TFT. TFT telah dipilih kerana keupayaannya untuk menangkap kebergantungan jangka panjang, menggabungkan kovariat statik dan kovariat berubah-ubah masa, dan menjana ramalan berasaskan kuantil yang boleh ditafsir. Sebelum pemodelan, saluran paip data yang komprehensif telah dibina untuk membersihkan, menggabungkan dan menyeragamkan peringkat negeri seperti kes yang disahkan, kemasukan ke hospital, dan lain-lain. Ciri terbitan seperti liputan vaksinasi, kadar penggunaan hospital dan cuti umum telah direka bentuk untuk meningkatkan kualiti ramalan. Strategi pembangunan termasuk ujian awal pelbagai model untuk menilai kesesuaian prestasi. Selepas penilaian, hanya model TFT telah dilaksanakan sepenuhnya kerana keupayaannya dalam peramalan siri masa pelbagai variasi. Sistem ini dibina menggunakan *PyTorch Forecasting* pada GPU NVIDIA A100. Saluran paip prapemprosesan data yang lengkap telah dibangunkan untuk membersihkan, mengubah dan menggabungkan data pelbagai sumber mentah termasuk kes yang disahkan, rekod kemasukan ke hospital, dan lain-lain manakala ciri utama kejuruteraan seperti liputan vaksinasi dan kadar penggunaan

hospital. Data yang diproses ini kemudiannya disimpan dalam pangkalan data NoSQL Firebase Firestore, membolehkan penyepaduan lancar antara prapemprosesan, latihan model dan penggunaan. Pada bahagian hadapan (*frontend*), antara muka web interaktif telah dibangunkan menggunakan React dan TypeScript untuk membentangkan cerapan model. Aplikasi ini menampilkan papan pemuka dinamik yang memaparkan ramalan kes COVID-19 7 hari lebih awal untuk setiap negeri di Malaysia. Yang penting, ramalan itu dibentangkan sebagai simulasi berdasarkan tetingkap ramalan sejarah dari 9 Oktober 2022 hingga 15 Oktober 2022, menggunakan data yang tersedia sehingga tarikh akhir 8 Oktober 2022. Simulasi ini mencerminkan keadaan inferens dunia sebenar untuk menilai prestasi model dalam amalan.

Kata Kunci: *Temporal Fusion Transformer*, Pembelajaran Mesin, ramalan wabak COVID-19 di peringkat negeri

Abstract

This project focuses on forecasting COVID-19 outbreaks at the state level in Malaysia by leveraging advanced machine learning techniques. Key aspects of the project include multi-source epidemiological data integration, time series modeling, and interpretability of model predictions. The central problem addressed is the difficulty in producing accurate, timely, and localized forecasts of new COVID-19 cases, especially as the pandemic transitions into an endemic phase. Challenges arise from data inconsistencies across sources, evolving policy impacts, and the need for dynamic modelling capable of handling complex temporal dependencies. To address this, the project proposes the development of a robust forecasting framework using the Temporal Fusion Transformer (TFT) model. TFT was selected for its ability to capture long-term dependencies, incorporate static and time-varying covariates, and generate interpretable quantile-based forecasts. Before modelling, a comprehensive data pipeline was constructed to clean, merge, and standardize state-level such as confirmed cases, hospital admissions, and etc. Derived features such as vaccination coverage, hospital utilization rate, and public holidays were engineered to enhance prediction quality.

The development strategy included initial testing of various models (e.g., XGBoost, LSTM) to assess performance suitability. After evaluation, only the TFT model was fully implemented due to its capabilities in multivariate time series forecasting. The system was built using PyTorch Forecasting on an NVIDIA A100 GPU. A robust data preprocessing pipeline was established to clean, transform, and merge raw multi source data including confirmed cases, hospitalization records, and etc while engineering key features such as vaccination coverage and hospital utilization rate. This processed data was then stored in a Firebase Firestore NoSQL database, enabling seamless integration between preprocessing, model training, and deployment. On the frontend, an interactive web interface was developed using React and TypeScript to present model insights. The application features a dynamic dashboard that displays 7-day-ahead COVID-19 case predictions for each Malaysian state. Importantly,

the prediction is presented as a simulation based on a historical forecasting window from 9 October 2022 to 15 October 2022, using data available up to the cutoff date of 8 October 2022. This simulation mirrors real-world inference conditions to evaluate how the model would have performed in practice. The results illustrate the model's ability to anticipate rising case trends and respond to temporal shifts in public health dynamics.

Keywords: Temporal Fusion Transformer, Machine Learning, State Level Covid-19 Prediction

1.0 PENGENALAN

Pemantauan wabak penyakit yang berkesan adalah penting untuk tindak balas segera dan usaha pembendungan strategik. Dengan memantau wabak secara teliti, pihak berkuasa kesihatan dapat mengenal pasti kawasan yang terjejas dengan pantas, menganalisis corak penyebaran, dan melaksanakan intervensi yang disasarkan untuk mengurangkan penyebaran selanjutnya. Pengesanan awal membolehkan sistem kesihatan awam mengagihkan sumber dengan lebih cekap dan menyediakan pelan tindak balas lebih awal. Pengenalpastian pantas juga membolehkan pelaksanaan langkah-langkah pencegahan, seperti kempen vaksinasi atau protokol kuarantin, yang dapat mengurangkan kesan terhadap kesihatan dan ekonomi.

Salah satu contoh yang jelas adalah penyebaran wabak MERS (*Middle East Respiratory Syndrome*) yang merupakan sejenis penyakit pernafasan. Ia disebabkan oleh virus koronavirus (MERS-CoV) yang tersebar dari haiwan ke manusia melalui hubungan fizikal langsung. Gejala MERS boleh menunjukkan gejala seperti demam, batuk, dan kesukaran bernafas. Beberapa kes telah dikesan di mana pesakit tidak menunjukkan gejala apa-apa. Semasa tahun 2015, Korea telah dilanda wabak MERS dengan mencatatkan 185 kes yang telah disahkan dan seramai 36 kes kematian disebabkan oleh wabak tersebut. Wabak tersebut bermula apabila seorang warganegara Korea pulang dari perjalanan perniagaan merentasi beberapa negara Timur Tengah. Pada masa itu, Korea Selatan tidak mempunyai model ramalan atau data berkaitan untuk menjangka wabak tersebut, yang menyebabkan kelewatan dalam langkah pencegahan terhadap wabak tersebut. Kejadian ini bukan sahaja menimbulkan risiko kesihatan awam yang ketara tetapi juga menyebabkan kerugian ekonomi yang besar, terutamanya dalam sektor pelancongan dan aktiviti sosial. MERS, yang mula dilaporkan di Arab Saudi pada tahun 2012, kemudiannya merebak ke Eropah dan seterusnya sampai ke Korea (HA Mohd et al., 2016).

Penyakit berjangkit amat berbeza mengikut ciri-cirinya dan boleh dikategorikan kepada empat jenis kategori utama: sporadik (kejadian yang jarang berlaku), endemik (kelaziman biasa dalam kawasan geografi tertentu), hiperendemik (tahap penyakit yang tinggi dan berterusan), dan epidemik (peningkatan kes penyakit yang mendadak dan tidak dijangka). Jika epidemik merebak merentasi beberapa negara atau benua, ia menjadi pandemik, seperti COVID-19.

Pada masa ini, kebanyakan model ramalan dilatih semata-mata pada kes yang disahkan, bergantung pada trend ini untuk membuat ramalan perkembangan masa hadapan. Walau bagaimanapun, kerumitan penyakit berjangkit adalah lebih kompleks. Untuk meningkatkan ketepatan ramalan, adalah penting untuk memasukkan faktor tambahan seperti kes kematian dan rekod kemasukan ke hospital. Faktor-faktor ini memberikan perspektif yang lebih komprehensif tentang dinamik penyakit dan meningkatkan kebolehpercayaan ramalan.

Banyak negara menghadapi kesukaran dalam melaporkan wabak berjangkit disebabkan oleh infrastruktur perubatan yang masih asas dan ekosistem sistem kesihatan awam yang tidak lengkap, yang memburukkan lagi situasi tersebut. Sebagai contoh, semasa wabak virus Zika 2015 di Brazil, pihak berkuasa tempatan menghadapi kesukaran untuk bertindak balas dengan berkesan dan segera kerana kekurangan maklumat tentang virus tersebut dan vektornya, yang menghadkan usaha global untuk mengawal penyebarannya ke rantau lain (J. Ikejezie et al., 2016). Cabaran sedemikian juga memberi kesan kepada kualiti model ramalan. Rangka kerja prapemprosesan data yang tidak konsisten atau kurang jelas boleh menjejaskan kualiti data yang digunakan untuk latihan model, secara langsung menjejaskan prestasi dan kebolehpercayaan model. Memastikan saluran paip prapemprosesan yang mantap adalah penting untuk mengurangkan isu ini dan meningkatkan ketepatan model.

Oleh itu, projek ini adalah satu inisiatif berasaskan data yang bertujuan untuk meningkatkan tahap kesihatan awam dan menlengkapkan ekosistem data berkaitan penyakit wabak berjangkit terutamanya Covid-19 dengan memanfaatkan pembelajaran mesin moden dan analitik data masa nyata. Dengan menganalisis pelbagai sumber data epidemiologi seperti kes COVID-19 yang disahkan, rekod kemasukan ke hospital ,data mobility,vaksinasi dan lain-lain, sistem ini akan memberikan ramalan yang lebih lengkap berdasarkan dinamik wabak penyakit tersebut. Ramalan ini dapat membantu pihak berkuasa kesihatan awam bertindak balas dengan cepat sebelum wabak penyakit tersebut merebak, berpotensi uttuk menyelamatkan nyawa dan mengoptimumkan pengagihan sumber secara efisien dan menggelakkan pembaziran sumber.

Projek ini bertujuan untuk merapatkan jurang antara kaedah pengawasan penyakit tradisional, yang sering bergantung kepada laporan kes yang tertangguh, dengan model ramalan masa nyata. Melalui integrasi data sejarah dan analitik masa nyata, sistem ini akan memberikan amaran awal untuk pihak berkuasa kesihatan dan orang awam, sekali gus membolehkan pendekatan yang lebih proaktif dalam pencegahan dan kawalan penyakit.

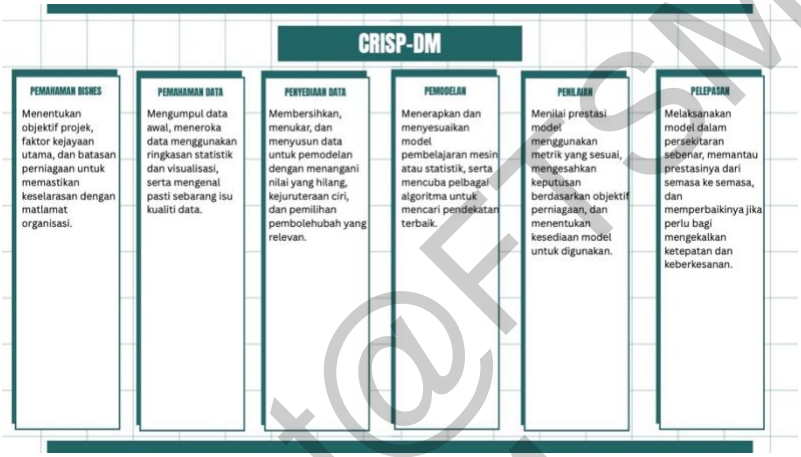
2.0 KAJIAN KESUSATERAAN

Ulasan literatur dan penilaian awal terhadap beberapa pendekatan pemodelan telah dijalankan dan terdapat beberapa model ramalan telah dikenal pasti sebagai calon berpotensi untuk latihan menggunakan data epidemiologi siri masa, termasuk kes yang disahkan COVID-19, rekod kematian, data mobiliti, dan data berkaitan lain. Model-model seperti XGBoost dan LSTM telah dipertimbangkan kerana keupayaan mereka dalam mengendalikan corak siri masa. Selain itu, model-model lanjutan seperti *Neural Relational Autoregression* (NRA), *Recurrent Neural Networks* (RNN) yang disepadukan dengan model SEIR, serta T*emporal Fusion Transformer* (TFT) turut dinilai berdasarkan kesesuaian mereka terhadap sifat data yang kompleks dan pelbagai sumber.

Namun begitu, berdasarkan hasil perbandingan prestasi dan keupayaan setiap model dalam menangani pemboleh ubah masa nyata dan struktur data multivariat, hanya model Temporal Fusion Transformer (TFT) telah dipilih untuk dibangunkan sepenuhnya. TFT menawarkan kelebihan seperti keupayaan perhatian (*attention mechanisms*), pengendalian pemboleh ubah statik dan dinamik, serta ramalan berdasarkan kuantil yang menjadikannya lebih sesuai dalam meramalkan trend wabak. Pendekatan ini memastikan bahawa model yang dipilih benar-benar sesuai dengan struktur data dan keperluan ramalan projek ini, sekaligus mengoptimumkan ketepatan dan kebolehpercayaan hasil ramalan.

3.0 METODOLOGI

Metodologi yang digunakan untuk kajian model ialah *Cross Industry Standard Process for Data Mining* (CRISP-DM). Metodologi ini adalah model proses yang berperanan sebagai asas untuk proses sains data. Ia mempunyai enam fasa berurutan iaitu Pemahaman Perniagaan (*Business Understanding*), Pengetahuan Data (*Data Understanding*), Penyediaan Data (*Data Preparation*), Pemodelan (*Modelling)*, Penilaian (*Evaluation*) dan Perlepasan (*Deployment*)



3.1 Fasa Pemahaman Data

Fasa ini bertujuan untuk memperoleh pemahaman mendalam tentang data yang tersedia, termasuk sumbernya, struktur, kandungan, dan hubungan antara pembolehubah, serta mengenal pasti potensi isu yang boleh menjejaskan pemodelan.

3.1.1 Objektif dan Peranan Data dalam Projek

Untuk memodelkan dinamik penyebaran COVID-19 secara berkesan, projek ini mengintegrasikan pelbagai set data heterogen dari sumber rasmi dan dipercayai. Setiap sumber menyumbang kepada pembentukan pemahaman yang menyeluruh terhadap faktor-faktor yang mempengaruhi wabak iaitu daripada bilangan kes kepada tingkah laku penduduk dan polisi kerajaan. Fokus utama adalah pada ramalan jangka pendek kes baharu yang disahkan, tetapi dengan mengambil kira intervensi struktur dan tingkah laku yang berubah dari semasa ke semasa.

3.1.2 Sumber Data Utama

Set data yang digunakan termasuk:

Sumber	Kandungan	Peranan dalam Pemodelan
MOH - Data Kes COVID-19	Kes baharu, pemulihan, kematian	Sasaran utama (cases_new_per100k_7d_avg)

MOH - Data Vaksinasi	Liputan vaksin penuh dan penggalak	Penanda aras imuniti populasi
MOH - Data Penghospitalan	Kemasukan, penggunaan ICU, kadar penggunaan katil	Indikator tekanan sistem kesihatan
Google Mobility	Perubahan pergerakan ke lokasi tertentu	Proksi tingkah laku sosial
Oxford OxCGRT	Indeks polisi kerajaan	Model intervensi dasar
***Our World in Data* (Rt)**	Kadar pembiakan virus	Penunjuk arah aliran jangkitan
Maklumat Masa Hadapan	is_weekend, is_holiday, holiday_type	Dimasukkan sebagai known future covariates

3.1.3 Struktur Data dan Skop Temporal

- Skop Geografi: Semua 15 negeri di Malaysia.
- Resolusi Masa: Harian, diselaraskan antara semua sumber.
- Tempoh Data: Merangkumi beberapa gelombang utama wabak, dengan struktur berterusan dari tarikh pertama hingga tarikh akhir dalam set data.

3.1.4 Pemerhatian Awal dan Isu Potensi

- Kehilangan Data: Ciri seperti vaksinasi dan penghospitalan mempunyai kekosongan nilai yang perlu diimputasi.
- Perbezaan Resolusi Masa: Sesetengah data, seperti kadar pembiakan (Rt), hanya tersedia di peringkat kebangsaan dan perlu dipetakan secara berhati-hati.
- Kepelbagaian Skala dan Unit: Ciri berangka merentasi skala yang berbeza, memerlukan penskalaan semasa fasa pemprosesan.

3.2 Fasa Penyediaan Data

Fasa ini melibatkan semua aktiviti transformasi data mentah ke dalam bentuk yang bersih, berstruktur, dan sesuai untuk pemodelan menggunakan *Temporal Fusion Transformer* (TFT).

3.2.1 Pengambilan Data

- Semua set data disimpan dalam format CSV di Firebase Firestore.
- Data diekstrak ke dalam persekitaran Google Colab dan dimuatkan ke dalam struktur Pandas DataFrames untuk memudahkan manipulasi.

3.2.2 Penjajaran Temporal

- Semua rekod diindeks semula kepada kekerapan harian dan diselaraskan mengikut tarikh dan negeri.
- Ini memastikan konsistensi antara sumber dan membolehkan model mengenal pasti kebergantungan masa.

3.2.3 Pembersihan Data

- Teknik imputasi digunakan untuk mengisi nilai hilang:
  - *forward-fill* untuk ciri bersiri seperti *vaccination_rate, icu_admissions*.
  - *zero-fill* untuk ciri binari seperti *is_holiday*.
- Medan yang tidak relevan dan duplikasi dipadamkan.

3.2.4 Kejuruteraan Ciri

- Ciri Ketinggalan dan *Rolling Average*:
  - Lag 1, 3, 7, dan 14 hari untuk *cases_new, vaccinations, hospitalizations*.
- Ciri Kadar:
  - *vaccination_coverage_full, hospital_utilization_rate*.
- Ciri Kategori Temporal:
  - *day_of_week, is_weekend, month*.
- Input Masa Depan yang Diketahui:
  - *is_holiday, holiday_type, school_closure*, digunakan sebagai *known future covariates* dalam TFT.

3.2.5 Normalisasi dan Penskalaan

- *StandardScaler* digunakan untuk penskalaan ciri berangka.
- Dilakukan berasingan untuk setiap negeri bagi mengekalkan perbezaan dinamik epidemik antara negeri.

3.2.6 Pengekodan Kategori

- Semua pembolehubah kategori ditukar kepada bentuk numerik menggunakan *Label Encoding*.
- Ini membolehkan integrasi dengan lapisan embedding dalam TFT.

3.2.7 Spesifikasi Input untuk TFT

Ciri-ciri dikategorikan untuk input TFT seperti berikut:

- *Static Categorical Variables: state*
- *Known Real Variables: stringency_index, vaccination_coverage*
- *Observed Real Variables: cases_new, hospitalizations, mobility*
- *Known Categorical Variables: day_of_week, is_weekend, is_holiday*
- *Time Index: time_idx* untuk urutan *temporal*

3.2.8 Penyimpanan dan Pembahagian Data

- Data akhir disimpan dalam struktur NoSQL di *Firebase Firestore*.
- Dataset dibahagikan kepada train, validation, dan test sets berdasarkan masa, memastikan tiada kebocoran maklumat ke masa hadapan semasa latihan.

3.3 Fasa Pemodelan *Temporal Fusion Transformer* (TFT)

3.3.1. Seni Bina Model TFT

Temporal Fusion Transformer (TFT) ialah model pembelajaran mendalam khas untuk peramalan siri masa pelbagai ufuk yang boleh ditafsir. Ia menggabungkan:

- Penyekod berulang (RNN/LSTM),
- Mekanisme perhatian (*attention mechanism*),
- Mekanisme pintu (*gating*),
- Lapisan embedding, dan
- Keluaran regresi kuantil (*quantile regression output layer*).

Dalam projek ini, TFT digunakan untuk meramalkan kes harian COVID-19 untuk 7 hari akan datang bagi setiap negeri di Malaysia. Model ini direalisasikan menggunakan:

- *PyTorch Forecasting* sebagai kerangka utama pembinaan model.
- *PyTorch Lightning* sebagai platform latihan, memastikan modulariti dan keupayaan pemprosesan GPU.
- Input: Ciri-ciri bersiri masa merentas 15 negeri, termasuk penunjuk epidemiologi, mobiliti, dasar awam, dan data statik.

3.3.2. Integrasi *Optuna* dan *Weights & Biases*

Bagi menala prestasi model secara sistematik, kerangka *Optuna* digunakan untuk penalaan hiperparameter automatik, dengan sokongan pemantauan melalui *Weights & Biases* (W&B).

Kelebihan gabungan:

- Pencarian Cekap (*Efficient Search*): Menggunakan *Tree-structured Parzen Estimator (TPE)* dan pemangkasan awal (*pruning*) untuk menghentikan percubaan tidak menjanjikan.
- Pemantauan Masa Nyata: W&B memvisualisasikan metrik model dan membolehkan perbandingan percubaan secara interaktif.
- Rekod Penuh Eksperimen: Kebolehulangan dijamin dengan penyimpanan kod, konfigurasi sistem, dan nilai hiperparameter.

3.3.3. Hiperparameter yang Ditala

Optuna mencuba pelbagai kombinasi hiperparameter seperti:

Hiperparameter	Julat / Nilai	Fungsi
learning_rate	[1e-4, 5e-3]	Saiz langkah kemas kini
hidden_size	[16, 32, 64]	Saiz lapisan tersembunyi RNN

hidden_continuous_size	[8, 16, 32]	Dimensi embedding untuk input berterusan
dropout	[0.2, 0.5]	Regularisasi untuk mengelakkan overfitting
attention_head_size	[1, 2]	Bilangan kepala perhatian
batch_size	[64, 128]	Saiz kelompok data untuk setiap lelaran
gradient_clip_val	[0.01, 0.5]	Had maksimum untuk nilai kecerunan

Kod pelaksanaan objective() dan study menggunakan API Optuna membolehkan pengurusan trial, pencatatan dan pemilihan konfigurasi terbaik secara automatik.

3.4 Fasa Penilaian dan Keputusan Model TFT

3.4.1. Metrik Penilaian

Untuk menilai model, empat metrik utama digunakan, semuanya dikira atas set pengesahan:

1. *Mean Absolute Error* (MAE) - Ukur purata magnitud ralat, tanpa mengira arah.
2. *Root Mean Squared Error* (RMSE) - Menghukum ralat besar secara lebih keras (kuasa dua).
3. *Symmetric Mean Absolute Percentage Error* (SMAPE) - Metrik berskala dan simetri terhadap ramalan berlebihan / kekurangan.
4. *Bias* - Tanda dan magnitud kecenderungan ramalan.

Model TFT dilatih dengan quantile loss, bukan MSE atau MAE konvensional. Ini membolehkan output dalam bentuk julat ramalan (seperti P10, P50, P90), penting dalam senario tidak pasti seperti kesihatan awam.

Setiap percubaan dilog menggunakan val_mae, val_rmse, val_smape, val_bias, dan val_quantile_loss sebagai objektif utama.

3.4.2. Keputusan Penalaan Optuna

Sebanyak 40 percubaan Optuna dijalankan. Setiap model:

- Dilatih selama 15 epoch sahaja,
- Dijalankan menggunakan GPU NVIDIA A100,
- Tujuan: eksplorasi tetapan hiperparameter bukan penumpuan akhir.

3.5 Fasa Perlepasan

Model akhir dilatih menggunakan semua data sehingga 8 Oktober 2022, tanpa set pengesahan, bagi menyamai senario dunia sebenar. Ramalan dijana untuk tempoh 9–15 Oktober 2022 merentas semua negeri dan Wilayah Persekutuan (kecuali Putrajaya).

4.0 HASIL KAJIAN

4.1 Prestasi dan Interpretasi Model

4.1.1 Prestasi Model Merentas Negeri

Model Temporal Fusion Transformer (TFT) menunjukkan variasi prestasi merentas negeri dalam menjangkakan kes COVID-19. Jadual berikut merangkumkan metrik utama seperti MAE, RMSE, SMAPE, dan Bias:

Negeri	MAE	RMSE	SMAPE	Bias
Selangor	0.5523	0.6368	3.13	+0.33
Kelantan	0.3424	0.5169	7.52	+0.30
Terengganu	0.3591	0.4182	16.59	-0.24
Pahang	0.6188	0.6728	25.91	-0.62
Perak	1.4281	1.4767	16.19	-1.43
W.P. Labuan	1.5087	1.5371	99.99	-1.51
Johor	1.7131	1.7730	41.37	+1.71
Sarawak	1.8372	1.8441	82.48	-1.84
Negeri Sembilan	1.9149	2.1108	17.14	-1.91
Pulau Pinang	2.2484	2.3585	37.78	-2.25
Kedah	2.3377	2.3474	82.91	-2.34
Melaka	2.5686	2.6378	19.86	+2.57
W.P. Kuala Lumpur	2.9859	3.1364	8.33	-2.99
Sabah	3.0029	3.0042	91.72	-3.00
Perlis	1.1083	1.1540	81.92	-1.11

4.1.2 Interpretasi Model dan Mekanisme Perhatian

Model TFT dilengkapi dengan mekanisme perhatian (attention mechanism) terbina dalam yang memberikan skor kepentingan terhadap ciri-ciri yang digunakan semasa ramalan. Ciri-ciri ini dipisahkan mengikut jenis: statik, pengekod (encoder) dan penyahkod (decoder).

a. Ciri Statik (Static Variables)

Ciri	Kepentingan
Population	4.0918

State	1.3411

Penjelasan:

- Population adalah penting kerana model meramalkan kadar per kapita. Saiz populasi negeri menjejaskan normalisasi kadar jangkitan dan kematian, menjadikannya kritikal untuk perbandingan silang negeri.
- State (ditunjukkan melalui embedding) memberikan konteks lokal yang membolehkan model mempelajari heterogeniti struktur sosial, amalan kesihatan awam, atau dasar antara negeri.

b. Ciri Pengekod (Encoder Variables)

Pembolehubah	Kepentingan Anggaran
daily_full_per100k_7d_avg	7.8354
cases_new_lag_3/7/14_7d_avg	~0.5
mobility (grocery, transit)	~0.5
deaths_new_per100k_7d_avg, vaccination_coverage_full	~0.3–0.4
time_idx, month, recovered_cases, is_weekend	<0.3

Penjelasan:

- `*daily_full_per100k_7d_avg*` adalah ciri paling penting kerana ia mencerminkan tahap imuniti komuniti semasa. Vaksinasi penuh memberi kesan langsung terhadap pengurangan kadar jangkitan dan kebolehjangkitan.
- Ciri `*cases_new_lag_3/7/14_7d_avg*` mencerminkan dinamika trend kes sejarah (autoregresif), membolehkan model mengesan pola peningkatan atau penurunan dalam masa dekat.
- Ciri-ciri mobiliti seperti `*grocery_and_pharmacy_percent_change*` dan `*transit_stations_percent_change*` menunjukkan sejauh mana aktiviti sosial dan pergerakan awam mempengaruhi kebarangkalian penyebaran penyakit.
- `*deaths_new_per100k_7d_avg*` dan `*hospital_utilization_rate*` berperanan sebagai penanda tahap keterukan wabak dan tekanan ke atas sistem kesihatan, yang mempengaruhi keputusan dasar.
- `*vaccination_coverage_full*` menunjukkan sejauh mana liputan perlindungan vaksin telah dicapai, menjadi faktor penting dalam menstabilkan trend jangkitan.

c. Ciri Penyahkod (Decoder Variables)

Pembolehubah	Kepentingan
vaccination_coverage_booster	2.8317

Stringency index	2.5968
time_idx	2.4622
is_holiday, reproduction rate, month	~1.3–1.7
is_weekend, holiday_type, day_of_week	~0.3–1.1

Penjelasan:

- `*vaccination_coverage_booster*` adalah ciri penyahkod paling dominan kerana ia mencerminkan jangkaan perlindungan masa hadapan, kritikal dalam mengurangkan penyebaran kes pada minggu berikutnya.
- `*Stringency index*` menunjukkan keketatan dasar kawalan (cth. sekatan pergerakan). Walaupun pada peringkat nasional, impaknya menyeluruh dan membantu model memahami persekitaran dasar semasa.
- `*Reproduction rate (Rt)*` membantu model memahami kebolehjangkitan semasa virus dalam komuniti.
- `*is_holiday*`, `*month*`, dan `*day_of_week*` menggambarkan corak tingkah laku sosial dan pergerakan semasa musim cuti, hujung minggu, dan perubahan bulan, yang boleh mencetuskan kluster atau perubahan mendadak dalam kes.
- `*time_idx*` adalah penanda masa global dan membantu model mengekalkan urutan temporal dalam penjadualan ramalan.

d. Penurunan Kepentingan Semasa Fasa Endemik

Beberapa ciri yang sebelum ini penting semasa pandemik menunjukkan pengurangan kepentingan dalam fasa endemik:

- Ciri seperti `*retail_and_recreation_percent_change*`, `*workplaces_percent_change*`, dan `*residential_percent_change*` menunjukkan pemberat rendah, mencerminkan normalisasi tingkah laku sosial.
- `*is_weekend*` dan `*holiday_type*` turut menurun kerana pengaruh mereka terhadap pergerakan dan jangkitan menjadi kurang signifikan apabila masyarakat telah menyesuaikan rutin dan SOP.
- Walaupun `*vaccination_coverage_booster*` masih penting, perhatian terhadapnya sedikit berkurang kerana liputan tinggi vaksinasi mengurangkan variasi merentas masa.

4.1.3 Kesimpulan Interpretasi

Model TFT bukan sahaja menghasilkan ramalan yang tepat untuk kebanyakan negeri, tetapi juga menunjukkan kebolehtafsiran yang tinggi melalui mekanisme perhatian. Ciri-ciri yang paling berpengaruh dapat dikenalpasti dan ditafsir berdasarkan prinsip epidemiologi dan dasar awam. Ini menjadikan TFT sesuai untuk digunakan dalam konteks dunia sebenar, terutama dalam membuat keputusan dasar kesihatan awam yang pantas dan bersasar.

4.2 Antara Muka Pengguna

4.2.1 Teknologi dan Tumpuan Reka Bentuk

Bahagian hadapan sistem ramalan dibangunkan sebagai aplikasi web interaktif berasaskan komponen, bertujuan menyampaikan maklumat ramalan dengan jelas, cepat dan boleh digunakan oleh umum serta pembuat dasar.

Teknologi Digunakan:

1. *React + TypeScript* – Rangka kerja utama untuk pembangunan UI berstruktur.
2. *Tailwind CSS* – Penggayaan berasaskan utiliti untuk reka bentuk responsif.
3. *ShadCN UI* – Komponen antaramuka moden dan boleh diakses.
4. *Firebase Firestore* – Penyimpanan data ramalan dan sejarah.
5. *Vite* – Alat binaan yang mempercepatkan pembangunan tempatan.
6. *React Router* – Pengurusan navigasi dan laluan dinamik.

4.2.2 Susun Atur Aplikasi Web

a. Laman Utama

- Menu interaktif untuk memilih negeri.
- Paparan graf ramalan 7 hari kes COVID-19 bagi setiap 100,000 penduduk.
- Petua alat dan kad nilai untuk memaparkan angka spesifik.
- Data dari koleksi tft_predictions di Firestore.

b. Paparan Data Lalu

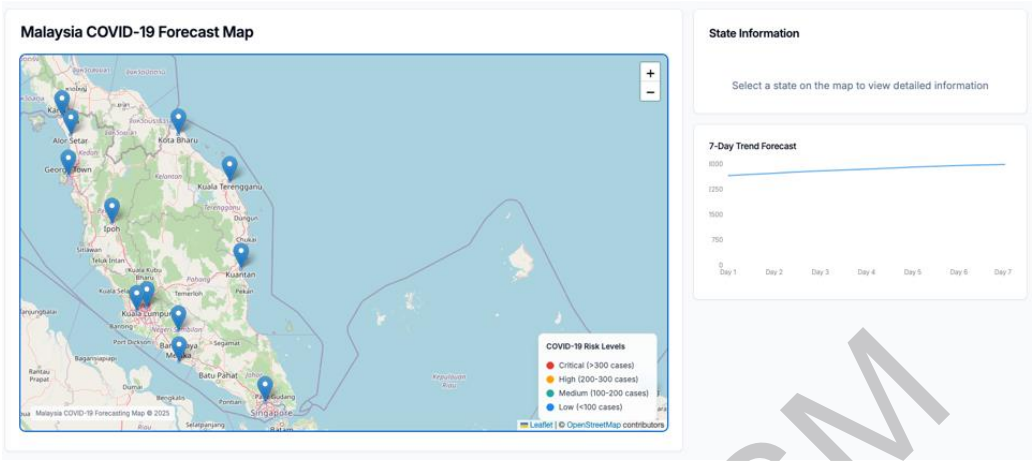
- Tunjuk data sejarah kes, kematian, penghospitalan, dan ICU.
- Penapis berdasarkan tarikh, negeri dan indikator.
- Diambil dari koleksi moh_cases, moh_deaths, dan moh_hospital.

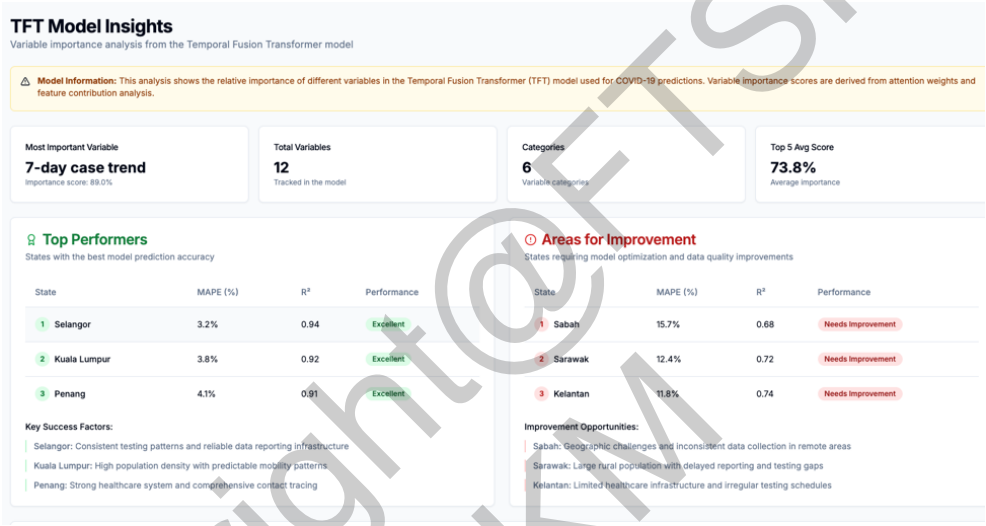
c. Papan Pemuka Model

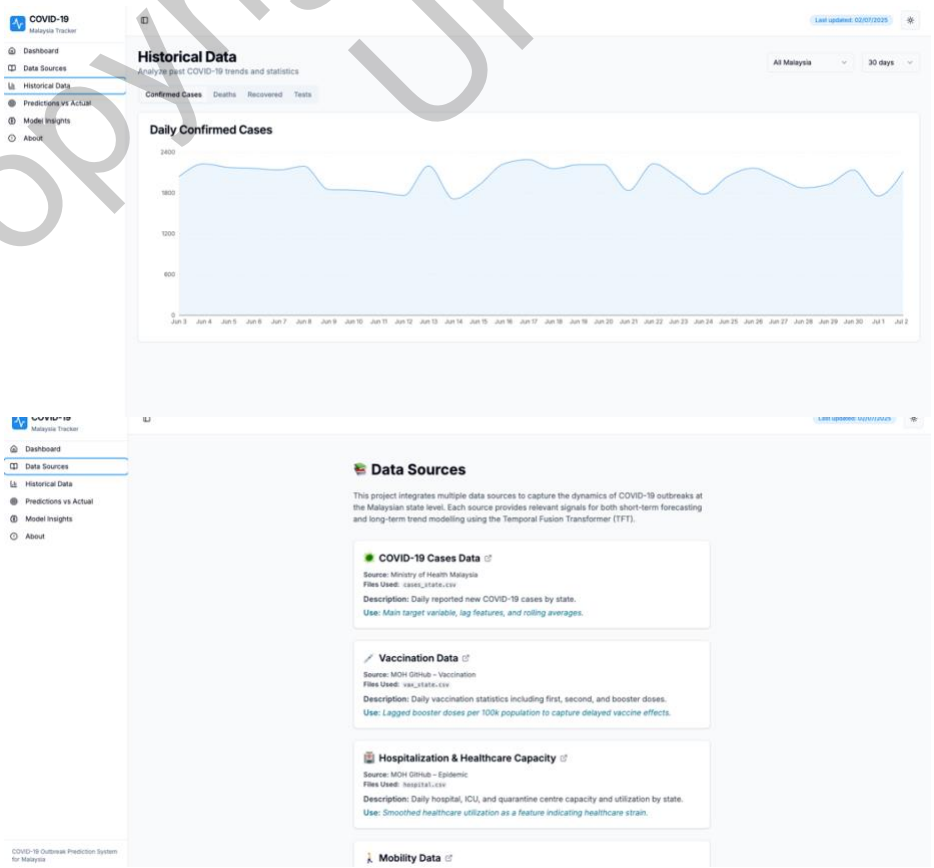
- Carta prestasi model: MAE, RMSE, SMAPE mengikut negeri.
- Bar chart dan jadual interaktif.
- Visualisasi kepentingan ciri mengikut pengekod dan penyahkod.
- Penjelasan tentang seni bina model dan batasannya.

d. Halaman Dokumentasi

- Maklumat projek, skop, sumber data, dan andaian.
- Senarai kredit, rujukan, dan latar belakang metodologi.







5.0 KESIMPULAN

Projek ini bertujuan untuk membangunkan sistem ramalan kes COVID-19 peringkat negeri di Malaysia menggunakan model pembelajaran mendalam *Temporal Fusion Transformer* (TFT), iaitu sebuah rangka kerja ramalan siri masa multivariat yang canggih dan boleh ditafsir. Sistem ini menggabungkan data epidemiologi yang telah dibersihkan dan diproses, ciri-ciri terbitan seperti liputan vaksinasi dan pemboleh ubah ketinggalan (*lag features*), serta pemboleh ubah masa hadapan seperti cuti umum dan indeks polisi kebangsaan. Sebuah aplikasi web turut dibangunkan untuk memaparkan ramalan dan analisis model secara interaktif kepada pengguna.

Kekuatan utama sistem ini terletak pada keupayaannya menangkap corak temporal yang kompleks merentasi negeri-negeri yang berbeza melalui satu model yang bersatu dan boleh ditafsir. Model TFT menunjukkan prestasi yang memberangsangkan, terutamanya dalam memahami trend sejarah dan mengintegrasikan pemboleh ubah masa hadapan yang telah diketahui. Mekanisme perhatian (*attention mechanism*) dalam TFT juga memberikan kejelasan tentang ciri-ciri yang paling mempengaruhi sesuatu ramalan.

Namun begitu, salah satu batasan penting ialah penggunaan pemboleh ubah dasar awam peringkat kebangsaan, terutamanya *Stringency Index* daripada *Oxford COVID-19 Government Response Tracker*. Walaupun pemboleh ubah ini dimasukkan secara eksperimen dan terbukti penting (kedua paling penting dalam dekoder model), ia kekurangan ketepatan dari segi butiran geografi. Beberapa negeri di Malaysia Timur, seperti Sabah dan Sarawak, menunjukkan ralat ramalan yang lebih tinggi, dipercayai berpunca daripada perbezaan dasar tempatan yang tidak diwakili dengan tepat oleh indeks kebangsaan tersebut.

Bagi meningkatkan ketepatan ramalan dan mengurangkan bias geografi, beberapa penambahbaikan disarankan iaitu menghasilkan atau mendapatkan data dasar awam peringkat negeri yang lebih terperinci, meneroka model berhierarki yang dapat membezakan antara kesan dasar nasional dan negeri, menggabungkan data mobiliti dan tingkah laku di peringkat negeri sebagai indikator tambahan, dan melanjutkan jangka masa simulasi untuk menguji ketahanan model dalam fasa wabak yang berbeza.

Kesimpulannya, projek ini berjaya menunjukkan potensi pembelajaran mendalam khususnya TFT dalam peramalan wabak secara tempatan, serta menekankan pentingnya penggunaan pemboleh ubah input yang relevan dari segi geografi. Iterasi masa hadapan boleh memperkukuh asas ini untuk menyokong perancangan kesihatan awam yang lebih tepat dan adil.

6.0 PENGHARGAAN

Dengan penuh rendah hati dan Syukur, saya TOW DING FENG, A192066, ingin merakamkan penghargaan ikhlas kepada semua pihak yang telah memberikan bantuan serta sokongan dalam penyediaan tesis ini. Terutamanya sekali, Syukur kepada Tuhan Yang Maha Esa, yang memberi petunjuk, semangat dan keberkatan sepanjang perjalanan penyelidikan ini, semoga segala usaha kita diberkati-Nya. Seterusnya, saya ingin Menyusun sepuluh jari mengucapkan terima kasih kepada penyelia saya, Prof. Madya Dr. Shahnorbanun Sahran, atas bimbingan dan segala nasihatnya sepanjang proses penyelidikan ini. Usaha dan pengorbanan beliau amat dihargai. Sumbangan dan kemudahan penyelidikan yang disediakan oleh Fakulti Teknologi dan Sains Maklumat Universiti Kebangsaan Malaysia juga sangat membantu dalam menjalankan kajian ini. Terima kasih atas sokongan infrastruktur dan kemudahan teknologi yang diberikan.

Di samping itu, saya ingin mengucapkan ribuan terima kasih kepada semua pihak terutamanya ibu bapa saya yang telah menyediakan sumbangan kewangan bagi kejayaan penyelidikan ini. Tanpa mereka, projek ini tidak akan menjadi suatu kenyataan. Ucapan terima kasih juga dihulurkan kepada setiap individu, rakan sejawat dan keluarga tersayang yang memberi motivasi dan sokongan moral sepanjang perjalanan kajian ini. Segala pengorbanan dan sokongan daripada setiap pihak telah membantu saya menghadapi segala cabaran dan halangan dalam menjayakan penyelidikan dan tesis ini. Semoga segala jasa baik ini diberkati. Terima kasih yang tidak terhingga kepada semua yang terlibat.

7.0 RUJUKAN

Amin, S., Uddin, M.I., AlSaeed, D.H., Khan, A., & Adnan, M. 2021. Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches. *Complexity* 2021: 1–12.

Eckerson, W. W. 2002. Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute*, 1–32.

Fadhil, M. 2024. 3.10 Penyediaan Spesifikasi Keperluan Sistem [F2.6]. Portal Jaminan Kualiti Perisian Aplikasi Sektor Awam UnIt Pemodenan Tadbiran Dan Pengurusan Malaysia (MAMPU). Retrieved from: https://sqa.mampu.gov.my/index.php/ms/3-10-penyediaan-spesifikasi-keperluan-sistem-f2-6

George, D., & Mallery, P. (2018). Descriptive statistics. In *IBM SPSS Statistics 25 Step by Step* (pp. 126–134). Routledge.

Inflectra. 2022. What Are System Requirements Specifications? Retrieved from: https://www.inflectra.com/Ideas/Topic/Requirements-Definition.aspx#:~:text=The%20primary%20reasons%20for%20their,factors%20that%20it%20must%20satisfy

Kim, J., & Ahn, I. 2021. Infectious disease outbreak prediction using media articles with machine learning models. *Scientific Reports* 11(1): 1–13.

Lim, B., Arik, S. O., Loeff, N., & Pfister, T. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. In *Proceedings of the Neural Information Processing Systems (NeurIPS 2021)*.

Marzouk, M., Elshaboury, N., Abdel-Latif, A., & Azab, S. 2021. Deep learning model for forecasting COVID-19 outbreak in Egypt. *Process Safety and Environmental Protection* 153: 363–375.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M.J., & Flach, P.A. 2021. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering* 33(8): 3048–3061.

Muhammad, L.J., Algehyne, E.A., Usman, S.S., Ahmad, A., Chakraborty, C., & Mohammed, I.A. 2021. Supervised Machine Learning Models for Prediction of COVID-19 Infection Using Epidemiology Dataset 2(1): 11.

Odu, N., Prasad, R., & Onime, C. 2021. Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked* 22: 100508.

Salim, N.A.M., Wah, Y.B., Reeves, C., Smith, M., Yaacob, W.F.W., Mudin, R.N., Dapari, R., Sapri, N.N.F.F., & Haque, U. 2021. Prediction of dengue outbreak in Selangor, Malaysia using machine learning techniques. *Scientific Reports* 11(1): 939.

SolveXia. 2023. Data Transformation Steps, Techniques & Tools. Retrieved from: https://www.solvexia.com/blog/data-transformation-steps

Song, Y., & Yoon, B. (2024). Prediction of infectious diseases using sentiment analysis on social media data. *PLoS ONE* 19(9): e0309842.

Understanding of LSTM Networks. (2020).

Understanding LSTM: Architecture, Pros and Cons, and Implementation. (2023).

Wathore, R., Rawlekar, S., Anjum, S.G., Gupta, A., Bherwani, H., Labhasetwar, N.K., & Kumar, R. 2022. Improving performance of deep learning predictive models for COVID-19 by incorporating environmental parameters. *Gondwana Research* 114: 69–77.

Wirth, R., & Hipp, J. 2000. CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39.

Zhang, T., Rabhi, F.A., Chen, X., Paik, H., & Macintyre, C.R. 2023. A machine learning-based universal outbreak risk prediction tool. *Computers in Biology and Medicine* 169: 107876.

Tow Ding Feng (A192066)

Prof. Madya Dr. Shahnorbanun Sahran

Fakulti Teknologi & Sains Maklumat

Universiti Kebangsaan Malaysia