Sains Malaysiana 54(7)(2025): 1785-1796
http://doi.org/10.17576/jsm-2025-5407-12

Meneroka Keberkesanan Diskriminan Fisher dalam Pengelasan Morfologi Galaksi

(Exploring the Viability of Fisher Discriminants in Galaxy Morphology Classification)

SAZATUL NADHILAH ZAKARIA, SANTTOSH MUNIYANDY & JOHN Y.H. SOO*

School of Physics, Universiti Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia

Diserahkan: 17 Mac 2025/Diterima: 19 Mei 2025

ABSTRAK

Salah satu cabaran terbesar dalam astronomi adalah pengelasan galaksi dengan tepat, terutamanya dalam membezakan antara jenis galaksi yang berbeza. Terdapat pelbagai algoritma kompleks yang telah menunjukkan prestasi tinggi dalam menjalankan tugas pengelasan, namun kerumitan algoritma ini kebiasaannya mengambil masa pemprosesan yang lebih lama dan sukar untuk difahami. Kajian kami menangani isu ini dengan meneroka keberkesanan diskriminan Fisher, suatu algoritma yang jauh lebih mudah dalam menjalankan pengelasan morfologi galaksi. Kami menguji empat algoritma pembelajaran mesin: diskriminan Fisher, Rangkaian Neural Buatan (ANN), Pokok Keputusan Tergalak (BDT) dan *k*-jiran terdekat (*k*NN) untuk mengelaskan galaksi berdasarkan bentuk bonjol pusat. Dengan menggunakan data dari Tinjauan Langit Digital Sloan (SDSS), kami menguji lima transformasi pemboleh ubah pra-pemprosesan: penormalan, nyahkorelasi, analisis komponen utama (PCA), penyeragaman dan Gaussanisasi, serta mengelaskan bentuk bonjol pusat galaksi kepada bentuk bulat atau tiada bonjol, berdasarkan Pokok Keputusan *Galaxy Zoo*. Apabila dibandingkan dengan label daripada *Galaxy Zoo* 2 (GZ2), diskriminan Fisher dengan transformasi penyeragaman memperoleh skor kejituan tertinggi iaitu 0.9310, melebihi ANN, BDT dan *k*NN masing-masing setinggi 1.93%, 0.42% dan 3.08%.

Kata kunci: Diskriminan Fisher; morfologi galaksi; pembelajaran mesin; pengelasan galaksi; rangkaian neural buatan

ABSTRACT

One of the major challenges in astronomy involves accurately classifying galaxies, particularly distinguishing between different galaxy types. While many complex algorithms have shown strong performance in classification tasks, their complexity often results in longer processing times and increased difficulty in understanding. This study addresses this issue by exploring the viability of Fisher discriminants, a much simpler algorithm, in performing galaxy morphology classification. We tested four machine learning algorithms: the Fisher discriminant, Artificial Neural Networks (ANNs), Boosted Decision Trees (BDTs), and *k*-Nearest Neighbours (*k*NNs) to classify galaxies by the shape of their central bulges. Using data from the Sloan Digital Sky Survey (SDSS), we utilised five pre-processing transformations: normalisation, decorrelation, principal component analysis (PCA), uniformisation, and Gaussanisation, and classified the shape of central bulge into either rounded or no-bulge, based on the Galaxy Zoo Decision Tree. When compared to the Galaxy Zoo 2 (GZ2) labels, the Fisher discriminant with uniformisation obtained the highest accuracy score of 0.9310, outperforming ANN, BDT, and kNN by 1.93%, 0.42%, and 3.08%, respectively.

Keywords: Artificial Neural Networks; fisher discriminants; galaxy classification; galaxy morphology; machine learning

PENGENALAN

Galaksi merupakan sebuah sistem yang terdiri daripada gas antara bintang, debu, bintang dan planet, yang mana semua elemen ini terikat oleh daya graviti. Secara amnya, pengelasan galaksi dilakukan mengikut urutan Hubble (Hubble 1926), iaitu sebuah rajah yang berbentuk tala bunyi yang membahagikan galaksi kepada galaksi pilin dan galaksi elips. Galaksi pilin terdiri daripada empat komponen: (1) sebuah kincir berputar gergasi yang mengandungi sistem lengan pilin berlingkar dan dipenuhi bintang-bintang muda, (2) sebuah cakera leper yang dipenuhi bintang-bintang tua, (3) bonjol pusat yang padat dengan bintang dan (4) halo yang mengelilingi galaksi, terdiri daripada bintang-bintang tua. Sementara itu, galaksi elips tidak mempunyai ciri dalaman nampak yang ketara dan bintang-bintang mengorbit terasnya secara rawak. Galaksi elips mengandungi kandungan gas dan debu yang rendah serta bintang-bintang yang lebih tua, maka kadar pembentukan bintang baharu adalah rendah. Galaksi elips dipercayai terbentuk hasilan daripada perlanggaran dan penyatuan galaksi pilin, oleh itu, ia kurang umum berbanding galaksi berpilin (Burkert & Naab 2003). Selain itu, terdapat juga galaksi kekanta, sejenis galaksi gabungan antara galaksi pilin dan galaksi elips yang mengandungi bonjol pusat dan sebuah cakera tanpa lengan pilin; dan juga galaksi tak nalar, iaitu galaksi yang tidak mempunyai

bentuk atau struktur yang tetap. Pengelasan galaksi berdasarkan sifatnya adalah penting untuk memahami dinamiknya demi memetakan struktur alam semesta (Buta 2011).

Pelbagai tinjauan langit telah dijalankan untuk memperoleh maklumat morfologi galaksi menerusi panjang gelombang cahaya yang berbeza, kemudiannya dianalisis menggunakan algoritma dan peralatan yang khusus. Sebagai contoh, Stoppa et al. (2023) telah membangunkan *AutoSourceID-Classifier* (ASID-C), suatu alat yang direka bentuk untuk menyelesaikan masalah pemisahan bintang-galaksi, menggunakan imej yang diambil oleh Mata Optik MeerKAT (MeerLICHT) (Paterson 2017), sebuah teleskop optik yang terletak di Sutherland, mengikut pengelasan morfologi Tinjauan Legasi Kamera Tenaga Gelap (DECaLS) (Dey et al. 2019). Hasil kajian mereka menunjukkan bahawa ASID-C secara tekal mengatasi prestasi pengelasan bintang-galaksi *SourceExtractor* (SE) (Bertin & Arnouts 1996), terutamanya apabila nisbah isyarat-ke-bunyi (S/N) pencerapan adalah rendah, menghasilkan ramalan yang lebih mantap dan dipercayai. Sementara itu, von Marttens et al. (2023) telah menggunakan pembelajaran mesin terselia bagi mengautomasikan *Tree-based Pipeline Optimization Tool* (TPOT) (Le et al. 2020) untuk menjalankan pengelasan bintang-galaksi-kuasar dalam sebuah sampel yang besar, seterusnya menghasilkan satu katalog nilai tambah (VAC) yang mengandungi pengelasan bagi 47.4 juta objek. Kajian mereka mendapati bahawa *eXtreme Gradient Boosting* (XGBoost) (Chen & Guestrin 2016) merupakan algoritma yang paling sesuai, mencapai kepersisan purata > 0.99 bagi galaksi dan bintang, serta > 0.96 bagi kuasar, mengatasi SE dan *stellar-galaxy loci classifier* (SGLC) (López-Sanjuan et al. 2019).

Salah satu set data yang terkemuka digunakan untuk mengesahkan pengelasan morfologi galaksi ialah *Galaxy Zoo* (Lintott et al. 2008), satu projek sains warga yang melibatkan sukarelawan dalam proses pengelasan galaksi berdasarkan bentuk dan ciri-cirinya. *Galaxy Zoo* 2 (GZ2) (Willett et al. 2013) merupakan kesinambungan daripada projek *Galaxy Zoo* yang menyediakan pengelasan morfologi yang lebih terperinci untuk sejumlah 304 122 galaksi SDSS yang paling besar dan terang. Urechiatu dan Frincu (2024) telah mengelaskan 10,000 imej galaksi daripada GZ2 kepada lima kelas: galaksi bulat dan halus, hampir halus, halus berbentuk cerut, berorientasi tepi dan berpilin, menggunakan model baharu berasaskan rangkaian neural perlingkaran (CNN). Mereka telah membandingkan keputusan model mereka dengan beberapa rangkaian neural seperti *DenseNet* (Huang et al. 2017), *EfficientNet* (Tan & Le 2019), *MobileNet* (Howard et al. 2017) dan kaedah Zhu et al. (2019), akhirnya menunjukkan bahawa model mereka mencapai pembaikan setinggi 1.7% dari segi kejituan dan kepersisan berbanding dengan model lain. Kini, model mereka menghasilkan anggaran keyakinan yang paling tinggi dalam pengelasan morfologi galaksi GZ2, yang bergantung sepenuhnya pada maklumat piksel imej. Data GZ2 turut digunakan dalam pelbagai penyelidikan pengelasan morfologi galaksi seperti Beck et al. (2018), Reza (2021) dan Vavilova et al. (2021), masing-masing menggunakan cara pengelasan, algoritma pembelajaran mesin dan strategi pensampelan yang berbeza. Ramai antaranya menekankan bahawa prestasi pengelas bergantung pada saiz dan keseimbangan set data.

Sejak kebelakangan ini, pembelajaran mesin telah menunjukkan kejayaan yang ketara dalam menyelesaikan masalah berkaitan astronomi dengan mengurangkan keperluan campur tangan manusia dan meningkatkan kecekapan. Berdasarkan kajian yang tersenarai, didapati bahawa kebanyakan kajian menggunakan algoritma yang kompleks dan diuji pada sampel bersaiz besar, mencecah jutaan titik data. Kaedah yang tersenarai ini mempunyai beberapa kelemahan seperti masa latihan yang lama, kesukaran pentafsiran hiperparameter dan penggunaan storan data yang tinggi. Sebaliknya, kaedah pembelajaran mesin yang lebih mudah seperti diskriminan Fisher (Fisher 1936) lebih cepat untuk dikendalikan, lebih mudah difahami dan memerlukan storan yang rendah, menjadikannya alternatif yang menarik. Diskriminan Fisher ialah suatu kaedah yang mengenal pasti gabungan linear antara parameter untuk memaksimumkan pemisahan kelas, dengan meminimumkan varians dalam-kelas sambil memaksimumkan varians antara-kelas. Raichoor et al. (2015) mendapati kaedah ini berkesan dalam memilih galaksi garis pancaran bagi tinjauan eBOSS, menyerlahkan keupayaannya dalam menyasarkan populasi galaksi tertentu dengan kejituan tinggi. Namun, Fraix-Burnet (2023) menyatakan bahawa kaedah ini mempunyai kelemahan apabila berhadapan dengan set data yang kompleks dalam pengelasan morfologi galaksi dan mencadangkan bahawa diskriminan yang lebih maju seperti rangkaian neural digunakan.

Pengujian prestasi diskriminan Fisher dalam mengelaskan galaksi dapat memberikan penemuan bernilai mengenai kecekapan dan potensinya, khususnya apabila dibandingkan dengan pendekatan semasa yang lebih kompleks. Oleh itu, dalam kajian ini, kami menyasarkan untuk mencapai objektif berikut: (1) menguji keberkesanan diskriminan Fisher dalam pengelasan morfologi galaksi dan membandingkan prestasinya dengan ANN, BDT dan kNN; serta (2) menganalisis kesan transformasi pemboleh ubah pra-pemprosesan terhadap diskriminan Fisher dan menilai ketepatannya.

Makalah ini disusun seperti berikut. Bahagian seterusnya membincangkan data dan metodologi yang digunakan. Konsep di sebalik perisian (ANNz2), set data (SDSS, GZ2) serta padanan magnitud dan parameter morfologi akan diterangkan. Seterusnya, empat algoritma pembelajaran mesin: diskriminan Fisher, ANN, BDT dan *k*NN akan dibincangkan, diikuti dengan pengenalan kepada lima transformasi pra-pemprosesan serta metrik yang digunakan untuk menilai prestasi algoritma tersebut.

Bahagian berikutnya membentangkan hasil kajian ini dan bahagian terakhir ialah rumusan kajian ini dan cadangan kajian masa hadapan.

BAHAN DAN KAEDAH

SET DATA

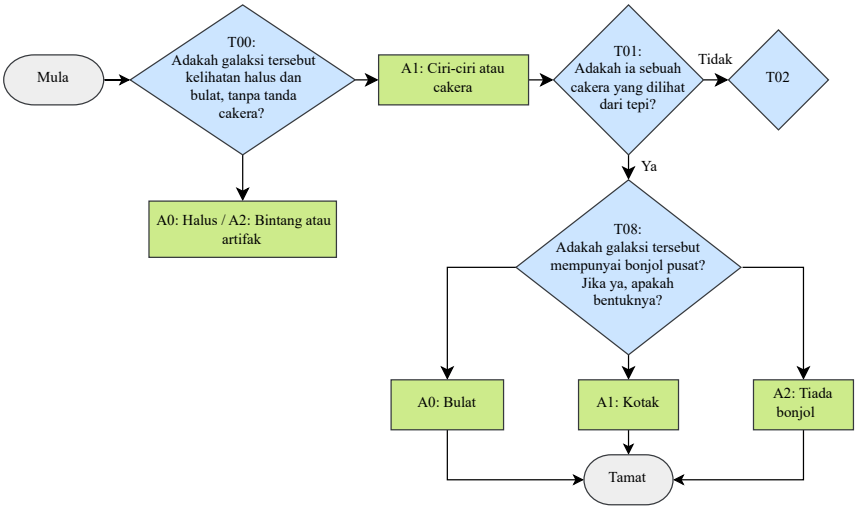
Kajian ini menggunakan data daripada Tinjauan Langit Digital Sloan (SDSS) dan *Galaxy Zoo* 2 (GZ2), dengan tumpuan khusus pada cabang yang menerangkan bentuk bonjol pusat dalam pokok keputusan GZ2, seperti yang diringkaskan dalam Rajah 1. SDSS merupakan salah satu tinjauan astronomi paling berimpak yang pernah dijalankan, ia bermula operasinya pada tahun 1998, dengan fasa pertamanya SDSS-I (York et al. 2000). Baru-baru ini, tinjauan ini telah melepaskan keluaran data (DR) ke-18 yang merupakan sebahagian daripada fasa kelimanya, SDSS-V (Almeida et al. 2023). SDSS menggunakan sistem fotometri *ugriz* (Doi et al. 2010; Fukugita et al. 1996), yang terdiri daripada 5 jalur lebar yang mencerap cahaya dalam panjang gelombang optik. Penggunaan jalur spektrum yang lebar ini memastikan kecekapan tinggi dalam pengesanan objek malap. Penapis tersebut disepadukan ke dalam peranti gandingan cas (CCD) dalam kamera fotometri SDSS (Gunn et al. 1998) yang merupakan salah satu komponen teleskop SDSS berdiameter 2.5 m (Gunn et al. 2006). Setakat ini, SDSS telah menyediakan data fotometri, morfologi dan spektroskopi bagi lebih daripada 200 juta objek samawi.

Sementara itu, GZ2 (Willett et al. 2013) merupakan salah satu daripada pelbagai projek sains warga yang dijalankan oleh *Zooniverse* (https://www.zooniverse.org), sebuah platform dalam talian yang membolehkan sukarelawan dari seluruh dunia menyertai kajian saintifik dalam pelbagai bidang. *Zooniverse* menggunakan intuisi dan pengalaman manusia untuk menambah kekompleksan dalam proses pengelasan objek, justeru, mempercepatkan penemuan saintifik dan meluaskan capaian data kajian kepada orang ramai. Secara ringkasnya, pengguna portal GZ2 akan ditunjukkan imej galaksi SDSS secara rawak dan diminta untuk mengelaskannya mengikuti set jawapan yang telah ditetapkan, berdasarkan 11 tugas pengelasan yang digariskan dalam Pokok Keputusan *Galaxy Zoo* (boleh didapati di https://data.galaxyzoo.org pada bahagian Data Visualizations). Projek ini berjaya tamat dalam tempoh 14 bulan dan menghasilkan set data yang merangkumi 16,340,298 pengelasan galaksi yang dilakukan oleh 83,943 sukarelawan (Willett et al. 2013). Selepas GZ2, projek *Galaxy Zoo* terus dilanjutkan kepada pelbagai set data pencerapan mahupun simulasi, menghasilkan tinjauan lain seperti *Galaxy Zoo* 3 *Hubble* (Willett et al. 2016), *Galaxy Zoo 4 Cosmic and Near-infrared Deep Extragalactic Legacy Survey* (GZ4 CANDELS) (Simmons et al. 2016) dan *Galaxy Zoo* 4 *Illustris* (Dickinson et al. 2018) dan semua telah tamat dengan jayanya.

Dalam kajian ini, kami memperoleh data pengelasan morfologi GZ2 (khususnya data yang mengelaskan bentuk bonjol pusat galaksi sebagai bulat (‘*rounded*’) atau tanpa bonjol (‘*no-bulge*’) dan memadankan sampel ini menggunakan ID objek tersebut dengan data fotometri daripada SDSS untuk menghasilkan set data yang mengandungi 1 530 galaksi. Data morfologi telah ditapis untuk mendapatkan sampel yang bersih iaitu `flag` = 1.

PEMBOLEH UBAH INPUT DAN SAIZ LATIHAN

Dalam setiap masalah pembelajaran mesin, pemilihan pemboleh ubah *input* (untuk melatih model) dan output (untuk diramalkan) memainkan peranan penting dalam menentukan keberkesanan model. Dalam kajian ini, kami menggunakan 11 parameter fotometri SDSS sebagai *input*, yang semuanya diperoleh daripada jalur *r* kerana ia mempunyai ketakpastian yang lebih rendah berbanding jalur fotometri lain. 11 parameter tersebut adalah seperti berikut:



RAJAH 1. Carta alir metodologi yang merumuskan cabang pokok keputusan *Galaxy Zoo* 2 yang digunakan dalam kajian ini

(a)Empat magnitud ketara, yang diukur berdasarkan fungsi sebaran titik (*point-spread function*) (`psfMag_r`), saiz gentian spektroskopi (`fiberMag_r`), padanan model galaksi (`modelMag_r`) dan padanan Petrosian (`petroMag_r`);

(b)Tiga jejari ketara Petrosian, iaitu jejari berkesan (*effective radius*, `petroRad_r`), jejari pada 50% cahaya (`petroR50_r`) dan jejari pada 90% cahaya (`petroR90_r`);

(c)Tiga kebarangkalian padanan model, iaitu log kebolehjadian (*log likelihood*) objek sebagai bintang (`lnLStar_r`), galaksi cakera bermodel eksponen (`lnLExp_r`) dan galaksi elips bermodel de Vaucouleurs (`lnLDeV_r`); dan

(d)Satu ukuran bentuk suai (*adaptive shape measure*, `mE1_r`), yang menerangkan keelipsan.

Perincian penuh parameter morfologi yang disebutkan boleh didapati dalam Stoughton et al. (2002). Kami memilih 11 parameter tersebut sebagai *input* untuk mengelaskan bentuk bonjol pusat galaksi berdasarkan andaian bahawa padanan model yang berbeza pada magnitud ketara, jejari dan keelipsan mungkin memainkan peranan penting dalam menentukan bentuknya. Sebagai contoh, galaksi dengan bonjol pusat yang bulat mungkin lebih berpadanan dengan model eksponen galaksi, sekali gus menunjukkan perbezaan ketara antara `lnLDeV_r` dan `lnLexp_r` serta perbezaan nilai jejari `PetroR50_r` dan `PetroR90_r`. Parameter output bagi kajian ini bernilai antara 0 dan 1, dengan nilai yang hampir kepada 1 menandakan galaksi yang bonjol pusatnya berbentuk bulat, manakala nilai yang hampir kepada 0 menandakan bahawa tiada bonjol dikesan. Kami menggunakan data pengelasan oleh orang awam menerusi GZ2 dalam kajian ini dan hanya data galaksi dengan ciri yang jelas dan terperinci diambil kira (`t09_bulge_shape_a25_rounded_flag` = 1 bagi galaksi bonjol pusat berbentuk bulat, `t09_bulge_shape_a27_no_bulge_flag` = 1 bagi galaksi tanpa bonjol).

Dengan penapisan data ini, kami memperoleh sebuah sampel sebanyak 1170 galaksi berbonjol pusat bulat dan 360 galaksi tanpa bonjol daripada SDSS DR8 dan jumlahnya 1530 galaksi. Berdasarkan Pokok Keputusan GZ2, terdapat satu lagi kategori iaitu galaksi berbentuk kotak (*boxy*) yang pada asalnya dikelaskan bersama-sama galaksi ini, namun ia dikecualikan daripada kajian kerana datanya tidak mencukupi (i.e., hanya terdapat 6 galaksi berbentuk kotak sahaja). Hal ini mungkin disebabkan galaksi daripada cabang tersebut kebanyakannya terdiri daripada jenis elips atau kekanta yang jarang dijumpai.

Untuk mengelakkan ketidakseimbangan bilangan galaksi dalam kelas masing-masing, kami hanya menggunakan 360 galaksi bagi setiap kategori, justeru, jumlah keseluruhannya adalah 720 galaksi. Set data ini dibahagikan dalam nisbah 2:2:1 untuk proses latihan, pengesahan dan penilaian. Set pengesahan digunakan untuk mengelakkan latihan lampau (*overtraining*), manakala set penilaian digunakan untuk menilai prestasi model. Isyarat dan latar belakang output ditentukan berdasarkan ambang 0.5 dengan kebarangkalian ≥ 0.5, dilabelkan sebagai isyarat (galaksi berbonjol pusat bulat) dan sebaliknya.

ANNz2: DISKRIMINAN FISHER DAN ALGORITMA LAIN

ANNz2[1] adalah suatu pakej perisian pembelajaran mesin yang menyediakan fungsi ketumpatan kebarangkalian (PDF) bagi masalah regresi dan pengelasan, bertumpuan pada aplikasi anjakan merah fotometri dan morfologi galaksi (Sadeh, Abdalla & Lahav 2016). ANNz2 dikodkan dalam *python* dan menggunakan pakej *Toolkit for Multivariate Data Analysis* (TMVA[2]) menerusi rangka kerja ROOT C++ (Brun & Rademakers 1997) untuk aplikasi pembelajaran mesin. ANNz2 mengandungi 5 konfigurasi: regresi tunggal, regresi rawak, pengelasan tunggal, pengelasan rawak dan pengelasan bin. Dalam kajian ini, versi ANNz2 2.3.2 telah digunakan bersama versi ROOT 6.30.02.

ANNz2 telah dipilih kerana ia membolehkan kami menjalankan kaedah diskriminan Fisher ke atas data kami, iaitu fokus kajian ini. Bagi menilai keberkesanannya, kami membandingkan prestasinya dengan tiga algoritma pembelajaran mesin lain, iaitu rangkaian neural buatan (ANN), pokok keputusan tergalak (BDT) dan *k*-jiran terdekat (*k*NN), yang menjadi penanda aras kami. Dalam perenggan berikutnya, empat algoritma pembelajaran mesin ini akan dihuraikan.

Analisis diskriminan secara amnya adalah suatu kaedah mencari gabungan linear parameter *input* bagi memisahkan objek ke dalam dua kelas atau ke atas. Dalam kes diskriminan Fisher (1936), andaikan x_k (*i*) ialah pemboleh ubah *input* ke-*k* bagi objek *i* dan y_F (*i*) ialah pengelasan output bagi objek *i*, jika sampel latihan kita mengandungi *Ns* objek isyarat (galaksi berbonjol pusat bulat) dan N_B objek latar belakang (galaksi tanpa bonjol), diskriminan Fisher akan menghasilkan pengelasan output

$$y_F(i) = F_0 + \sum_{k=1}^{n_{\mathrm{var}}} F_k x_k(i), \qquad (1)$$

dengan F_K ialah pekali Fisher, satu untuk setiap daripada $n_{\mathbf{var}}$ pemboleh ubah *input*; dan F_0 ialah pemalar yang memusatkan min sampel $\bar{y}_F$ bagi semua objek latihan *Ns* + N_B pada nilai sifar. Pekali Fisher bagi setiap pemboleh ubah *k* boleh ditulis sebagai

$$F_k = \frac{\sqrt{N_S N_B}}{N_S + N_B} \sum_{l=1}^{n} W_{kl}^{-1} \left(\bar{x}_{S,l} - \bar{x}_{B,l} \right), \qquad (2)$$

[1] Algoritma ini boleh dimuat turun dari https://github.com/IftachSadeh/ANNZ.

[2] Rujukan lanjut di https://root.cern/manual/tmva/.

yang berbentuk jumlah pendaraban antara perbezaan min sampel bagi isyarat dan latar belakang, serta songsang kepada matriks dalam-kelas (*within-class*) W_{kl}, dengan

$$W_{kl} = C_{S,kl} + C_{B,kl} = \sum_{k,l}^{N_S} (x_{S,k} - \bar{x}_{S,k})(x_{S,l} - \bar{x}_{S,l}) + \sum_{k,l}^{N_B} (x_{B,k} - \bar{x}_{B,k})(x_{B,l} - \bar{x}_{B,l}), \quad (3)$$

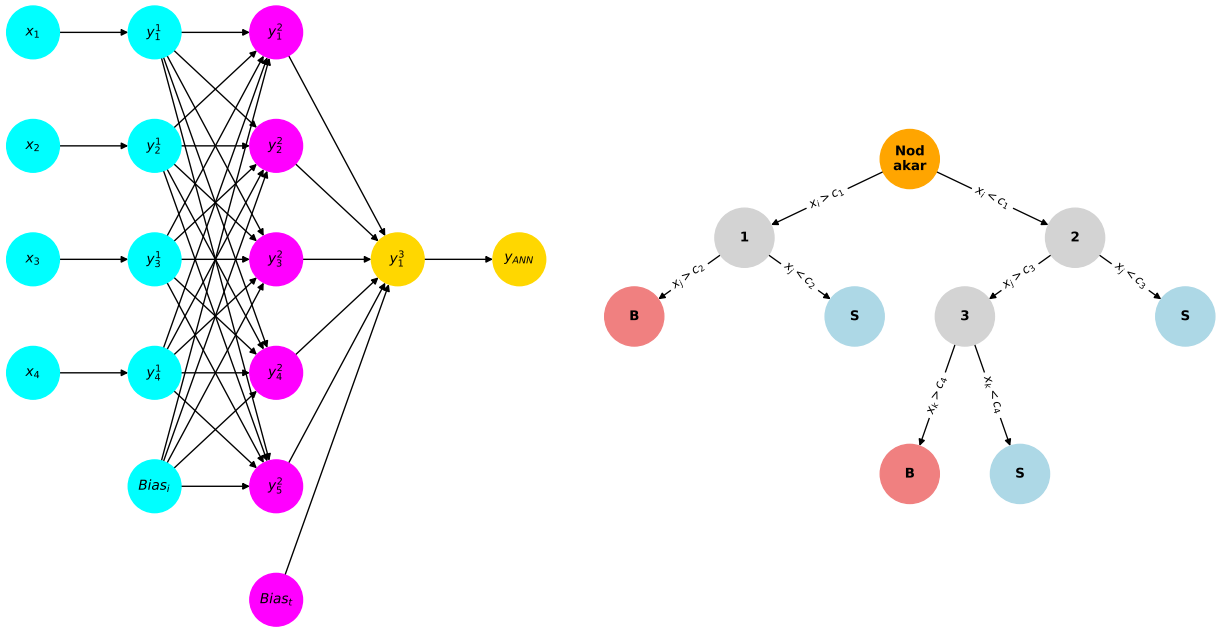
yang menerangkan sebaran peristiwa relatif kepada min bagi kelas masing-masing, iaitu $\bar{x}_S$ dan $\bar{x}_B$. Secara asasnya, $C_{S,kl}$ dan $C_{B,kl}$ hanyalah matriks kovarians bagi pemboleh ubah *input* dalam kelas isyarat atau latar belakang masing-masing.

Kaedah diskriminan Fisher meminimumkan sebaran dalam-kelas dengan output daripada kelas yang sama dikurung dalam sekitaran dekat sementara output daripada kelas yang berbeza dipisahkan sejauh mungkin antara satu sama lain. Walaupun kaedah ini ringkas, diskriminan Fisher berfungsi dengan baik, terutamanya bagi pemboleh ubah *input* bertaburan Gaussan yang berkorelasi linear; namun, kaedah ini akan gagal jika pemboleh ubah *input* mempunyai min sampel yang sama bagi isyarat dan latar belakang, justeru, transformasi pemboleh ubah ialah langkah pra-pemprosesan yang amat penting untuk pengelas ini.

ANN menganalisis dan memproses maklumat dengan cara yang menyerupai fungsi neuron dalam otak manusia (Reza 2021). Secara amnya, ia mengira hasil tambah berpemberat (*weighted sum*) bagi koleksi fungsi tindak balas antara parameter *input* dan output (Sadeh, Abdalla & Lahav 2016). ANN yang digunakan dalam ANNz2 berstruktur perseptron berbilang lapis (MLP) dengan perseptron disusun dalam sekurang-kurangnya tiga lapisan, iaitu lapisan *input*, terlindung dan output (Rajah 2). Dalam kajian ini, kami telah membina sebuah ANN ringkas dengan dua lapisan terlindung yang masing-masing mengandungi 8 dan 4 nod, setiap nod menggunakan fungsi pengaktifan tanh dan kaedah *Broyden-Fletcher-Goldfarb-Shannon* (BFGS) (Xie, Byrd & Nocedal 2020) digunakan bagi rambatan belakang. Keputusan purata diambil berdasarkan empat jalanan dengan nilai benih rawak yang berbeza. Untuk perincian konfigurasi ini, pembaca boleh merujuk kepada Hoecker et al. (2007).

BDT adalah sebuah pokok perduaan dengan keputusan bermula dari nod akar, sampel mengikuti satu siri pemecahan perduaan di nod dalaman, mengikuti cabang yang sepadan dengannya sehingga nod daun dicapai dan dipadankan dengan satu label (Reza 2021). Dalam proses pengelasan, keputusan sebuah daun biasanya dipecahkan kepada dua cabang, iaitu cabang isyarat dan cabang latar belakang (Rajah 2). Dalam kajian ini, kami menggunakan konfigurasi 500 pokok setiap hutan, kaedah galak *bagging* dan sekurang-kurangnya 2% sampel latihan diperlukan dalam setiap nod daun. Keputusan purata diambil daripada empat jalanan dengan konfigurasi pecahan nod yang berbeza.

*k*NN beroperasi berdasarkan prinsip kedekatan dengan peristiwa yang serupa lebih cenderung wujud berdekatan



RAJAH 2. Struktur perseptron lapisan berbilang (MLP, kiri) dan pokok keputusan tergalak (BDT, kanan). Rajah ini dihasilkan semula daripada Hoecker et al. (2007)

antara satu sama lain. Setiap sampel yang mengandungi *n* parameter *input* akan diplot dalam ruang *n*-dimensi dengan ciri *k*-jiran terdekatnya akan dipertimbangkan dalam ramalan parameter output. Jarak d_k diukur menggunakan fungsi metrik yang dipanggil jarak Euclidean,

$$d_k = \sqrt{\sum_{i=1}^{n} \left| x_i - y_{i,k} \right|^2}, \quad (4)$$

dengan x_i dan $y_{i,k}$ adalah titik data daripada sampel ujian dan sampel latihan *k*-terdekat, bagi *n* parameter *input*. Maka, pengelasan output adalah berdasarkan purata label *k*-jiran terdekat dengan *k* adalah hiperparameter yang dipilih oleh pengguna. Dalam kajian ini, kami menggunakan konfigurasi *k*NN yang berteras Gaussan dan berskala pecahan 90%. Keputusan diambil daripada purata empat jalanan dengan nilai *k* yang berbeza.

TRANSFORMASI PEMBOLEH UBAH PRA-PEMPROSESAN

Parameter *input* boleh dipraproses untuk tiga tujuan, iaitu mengurangkan korelasi antara parameter, menstabilkan varians pemboleh ubah dan mengurangkan masa tindak balas algoritma pembelajaran mesin (Hoecker et al. 2007). TMVA membenarkan penggunaan lima jenis transformasi pemboleh ubah pada parameter *input* sebelum proses latihan, iaitu

(1)**Penormalan**: penskalaan pemboleh ubah kepada nilai antara 0 hingga 1. Penormalan mengurangkan sisihan piawai pemboleh ubah dalam ruang *input*, tetapi tidak sesuai untuk sampel yang mengandungi banyak data pesisih;

(2)**Nyahkorelasi**: penghasilan pemboleh ubah yang tidak berkorelasi secara linear antara satu sama lain. Nyahkorelasi hanya berfungsi dengan baik dalam dua keadaan, iaitu jika pemboleh ubah pada asalnya berkorelasi secara linear atau bertaburan Gaussan. Jika tidak, ia akan meningkatkan tahap ketaklinearan antara pemboleh ubah, menjadikan proses latihan tidak berkesan. Nyahkorelasi membolehkan setiap parameter menyumbang secara tak bersandar, sekali gus mengurangkan penyuaian lampau (*overfitting*) dalam model;

(3)**Analisis Komponen Utama** (PCA): pengurangan dimensi pemboleh ubah bagi set data yang kompleks dan besar dengan penumpuan kepada komponen utama terawal (atau nilai eigen terbesar). PCA berfungsi dengan baik pada data yang mempunyai banyak parameter *input* kerana ia mengurangkan masa pengkomputeran di samping mengekalkan ciri data penting, supaya model memproses data berkorelasi lemah dan mencegah masalah penyuaian lampau;

(4)**Penyeragaman**: penggunaan fungsi taburan longgokan (CDF) yang diperoleh daripada data latihan untuk menyeragamkan taburan parameter; dan

(5)**Gaussanisasi**: transformasi pemboleh ubah *input* kepada taburan Gaussan dengan purata 0 dan sisihan piawai 1. Gaussanisasi mengurangkan kesan data pesisih semasa latihan.

Dalam kajian ini, selain daripada perbandingan prestasi latihan antara diskriminan Fisher, ANN, BDT dan kNN, kami juga menganalisis kesan penormalan (N), nyahkorelasi (D), PCA (P), penyeragaman (U) dan Gaussanisasi (G) terhadap setiap pengelas, berbanding dengan hasil tanpa sebarang transformasi pemboleh ubah (X).

METRIK PRESTASI

Metrik prestasi ialah ukuran yang digunakan untuk menilai keberkesanan pengelasan setiap kaedah pembelajaran mesin yang digunakan. Dalam kajian ini, kami menggunakan matriks kekalutan dan anggaran ketumpatan inti (KDE) Gaussan untuk memilih kaedah yang berprestasi keseluruhan paling baik.

Matriks kekalutan membandingkan hasil sebenar dengan hasil ramalan secara visual. Dalam masalah pengelasan perduaan, matriks kekalutan adalah matriks 2×2,

$$\begin{bmatrix} \text{Positif Benar (TP)} & \text{Negatif Palsu (FN)} \\ \text{Positif Palsu (FP)} & \text{Negatif Benar (TN)} \end{bmatrix}. \quad (5)$$

Menerusi matriks kekalutan, banyak metrik prestasi boleh dijana seperti kejituan (*accuracy*), kepersisan (*precision*), kepekaan (*recall*/*sensitivity*) dan skor F1, yang masing-masing berjulat 0 hingga 1 (De Diego et al. 2022). Kejituan digunakan untuk menilai kadar keseluruhan kes positif benar dan negatif benar,

$$\text{Kejituan} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{T}}{\text{P} + \text{N}}, \quad (6)$$

dengan T = TP + TN (jumlah benar), P = TP + FP (jumlah positif) dan N = TN + FN (jumlah negatif). Kepersisan mengukur ketepatan model dalam mengelaskan ramalan positif bagi kes positif benar dan positif palsu,

$$\text{Kepersisan} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (7)$$

Kepekaan menjamin sensitiviti model dalam mengenal pasti kejadian positif untuk menentukan sama ada model dapat mengelaskan isyarat dengan lebih baik berbanding latar belakang,

$$\text{Kepekaan} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (8)$$

Akhir sekali, skor F1 yang tinggi memberi jaminan bahawa model berfungsi dengan baik dari segi perkadaran dan keseimbangan antara kepersisan dan kepekaan,

$$\text{Skor F1} = 2 \times \frac{\text{kepersisan} \times \text{kepekaan}}{\text{kepersisan} + \text{kepekaan}}. \quad (9)$$

Selain itu, KDE Gaussan juga digunakan untuk menilai prestasi setiap kaedah pembelajaran mesin dengan memberi gambaran taburan data mengikut kelas yang diramalkan dan menghasilkan satu lengkung halus bagi PDF keseluruhan data. Di sini, kita dapat memahami ciri, kadar kecenderungan dan keberkesanan setiap kaedah pembelajaran mesin semasa mengelaskan kes positif dan negatif dengan memaparkan sebaran, ketumpatan dan pertindihan lengkung. Luas pertindihan lengkung yang tinggi bermaksud kaedah tersebut tidak dapat membezakan isyarat dan latar belakang secara berkesan. Dalam kajian ini, ambang keputusan bernilai 0.5 telah ditetapkan, diwakili oleh satu garis menegak yang menetapkan had pemisahan antara kelas dalam rajah KDE Gaussan (Weglarczyk 2018).

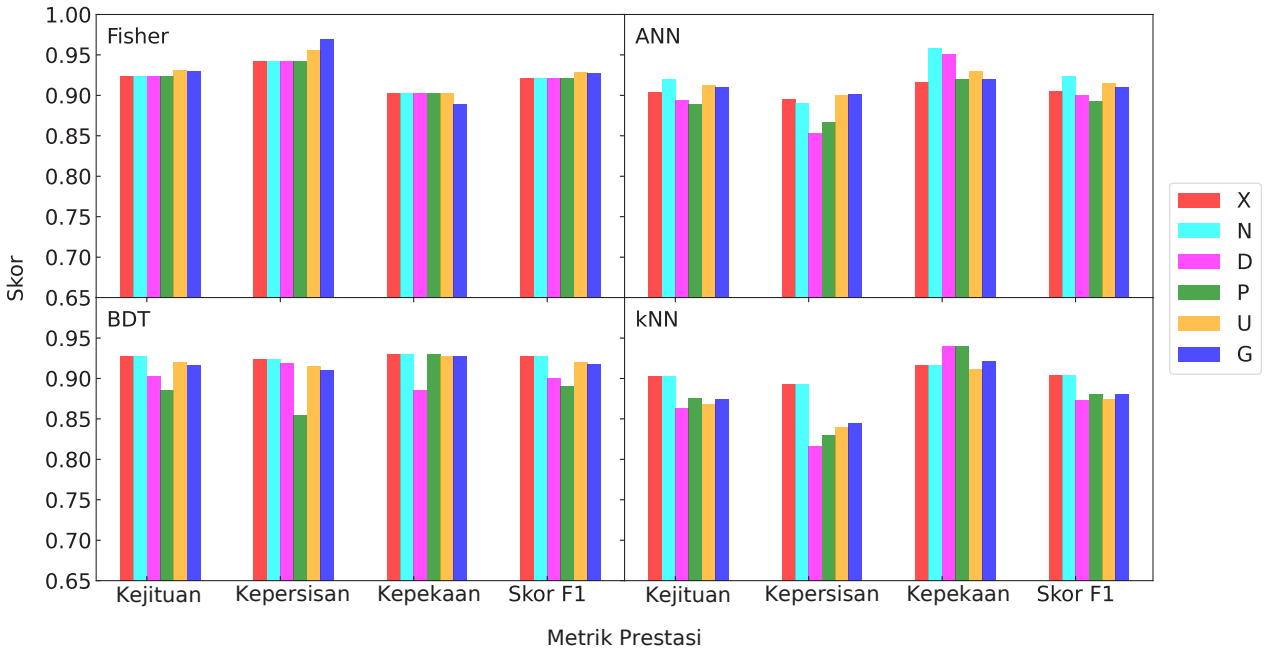
HASIL DAN PERBINCANGAN

KESAN TRANSFORMASI PEMBOLEH UBAH

Dengan membandingkan metrik prestasi daripada 24 (6 transformasi pemboleh ubah × 4 algoritma pembelajaran mesin) (Rajah 3), kami mendapati bahawa transformasi pemboleh ubah memberi kesan yang berbeza terhadap prestasi algoritma yang diuji. Keputusan ini dirumuskan dalam Jadual 1. Bagi diskriminan Fisher, transformasi penyeragaman dan Gaussanisasi meningkatkan kadar kejituan, masing-masing dengan perbezaan 0.80% dan 0.76% berbanding jalanan lalai (tanpa transformasi pemboleh ubah). Kesan yang sama dapat dilihat pada kepersisan (peningkatan sebanyak 1.48% dan 2.90%) dan skor F1 (peningkatan sebanyak 0.76% dan 0.59%). Walau bagaimanapun, Gaussanisasi menurunkan skor kepekaan sebanyak 1.55%.

Bagi ANN, penyeragaman dan Gaussanisasi meningkatkan prestasinya berbanding jalanan lalai dengan peningkatan kejituan sebanyak 0.96% dan 0.57%, kepersisan sebanyak 0.53% dan 0.66%, kepekaan sebanyak 1.5% dan 0.38%, serta skor F1 sebanyak 1.01% dan 0.54%. Penormalan juga menunjukkan peningkatan ketara dalam kejituan, kepekaan dan skor F1, masing-masing sebanyak 1.71%, 4.44% dan 1.91%. Bagi dua transformasi yang selebihnya, iaitu nyahkorelasi dan PCA, hanya kepekaan yang meningkat, masing-masing dengan perbezaan 3.72% dan 0.38% berbanding jalanan lalai.

Walaupun transformasi pemboleh ubah membawa kesan positif kepada Fisher dan ANN, perkara yang sama tidak berlaku bagi BDT dan *k*NN. Untuk BDT, semua transformasi menyebabkan penurunan skor selain



RAJAH 3. Perbandingan skor F1 bagi 5 transformasi pemboleh ubah pra-pemprosesan; tiada transformasi (X), penormalan (N), nyahkorelasi (D), analisis komponen utama (P), penyeragaman (U) dan Gaussanisasi (G), merentasi metrik prestasi untuk diskriminan Fisher (atas kiri), rangkaian neural buatan (ANN, atas kanan), pokok keputusan tergalak (BDT, bawah kiri) dan *k*-jiran terdekat (kNN, bawah kanan). Diperlihatkan bahawa prestasi diskriminan Fisher adalah paling baik dalam kejituan, kepersisan dan skor F1

JADUAL 1. Metrik prestasi dan ketakpastiannya (σ) bagi 6 transformasi pra-pemprosesan (X, N, D, P, U, G), untuk 4 algoritma pembelajaran mesin: diskriminan Fisher, rangkaian neural buatan (ANN), pokok keputusan tergalak (BDT) dan k-jiran terdekat (kNN). Nilai berwarna hijau dan biru masing-masing menandakan nilai tertinggi dalam setiap transformasi pemboleh ubah dan setiap metrik

Algoritma	Kejituan	σ	Kepersisan	σ	Kepekaan	σ	Skor F1	σ
X (Tiada Transformasi / Jalanan Lalai)								
Fisher	0.9236		0.9420		0.9028		0.9220	
ANN	0.9045	0.0077	0.8958	0.0145	0.9167	0.0098	0.9058	0.0071
BDT	0.9271	0.0045	0.9244	0.0082	0.9306	0.0000	0.9274	0.0041
kNN	0.9027	0.0106	0.8927	0.0188	0.9166	0.0000	0.9043	0.0095
N (Penormalan)								
Fisher	0.9236		0.9420		0.9028		0.9220	
ANN	0.9201	0.0083	0.8910	0.0141	0.9583	0.0000	0.9233	0.0075
BDT	0.9271	0.0045	0.9244	0.0082	0.9306	0.0000	0.9274	0.0041
kNN	0.9027	0.0106	0.8927	0.0188	0.9166	0.0000	0.9043	0.0095
D (Nyahkorelasi)								
Fisher	0.9236		0.9420		0.9028		0.9220	
ANN	0.8941	0.0017	0.8537	0.0046	0.9514	0.0040	0.8998	0.0013
BDT	0.9028	0.0063	0.9184	0.0130	0.8854	0.0183	0.9009	0.0071
kNN	0.8635	0.0101	0.8158	0.0141	0.9398	0.0046	0.8732	0.0083
P (Analisis Komponen Utama, PCA)								
Fisher	0.9236		0.9420		0.9028		0.9220	
ANN	0.8889	0.0085	0.8664	0.0119	0.9201	0.0066	0.8924	0.0078
BDT	0.8854	0.0134	0.8550	0.0204	0.9306	0.0000	0.8908	0.0113
kNN	0.8760	0.0096	0.8296	0.0195	0.9398	0.0046	0.8810	0.0103
U (Penyeragaman)								
Fisher	0.9310		0.9560		0.9030		0.9290	
ANN	0.9132	0.0083	0.9006	0.0175	0.9306	0.0057	0.9149	0.0071
BDT	0.9201	0.0035	0.9146	0.0076	0.9271	0.0035	0.9207	0.0031
kNN	0.8681	0.0120	0.8392	0.0185	0.9120	0.0046	0.8738	0.0101
G (Gaussanisasi)								
Fisher	0.9306		0.9697		0.8888		0.9275	
ANN	0.9097	0.0049	0.9018	0.0086	0.9201	0.0104	0.9106	0.0050
BDT	0.9167	0.0049	0.9097	0.0173	0.9271	0.0104	0.9177	0.0037
kNN	0.8750	0.0120	0.8445	0.0210	0.9213	0.0046	0.8808	0.0097

penormalan yang mengekalkan skor. Untuk *k*NN, hanya nilai kepekaan mengalami peningkatan berbanding jalanan lalai dengan peningkatan 2.5% bagi nyahkorelasi dan PCA, manakala Gaussanisasi meningkatkan kepekaan sebanyak 0.5%.

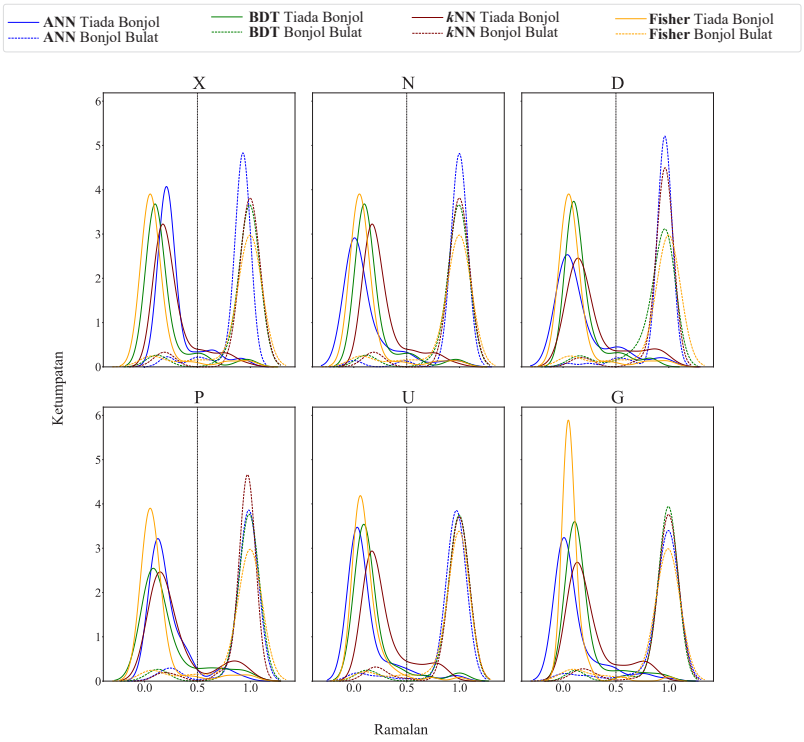
Oleh itu, kami menyimpulkan bahawa kesan transformasi pemboleh ubah pra-pemprosesan terhadap prestasi algoritma yang diuji berbeza secara signifikan bergantung pada setiap kaedah pembelajaran mesin. Diskriminan Fisher dan ANN mendapat manfaat paling besar daripada penyeragaman dan Gaussanisasi dengan peningkatan dalam ketepatan, *precision* dan skor F1. Sebaliknya, BDT dan *k*NN sama ada tidak mengalami perubahan atau penurunan dalam metrik prestasi dengan kebanyakan transformasi. Penemuan ini menekankan kepentingan memilih teknik pra-pemprosesan yang sesuai berdasarkan algoritma tertentu, menunjukkan bahawa pra-pemprosesan data adalah keperluan bagi diskriminan Fisher dan ANN, manakala proses ini boleh diabaikan apabila menggunakan BDT dan *k*NN.

PERBANDINGAN DISKRIMINAN FISHER DENGAN ANN, BDT DAN *k*NN

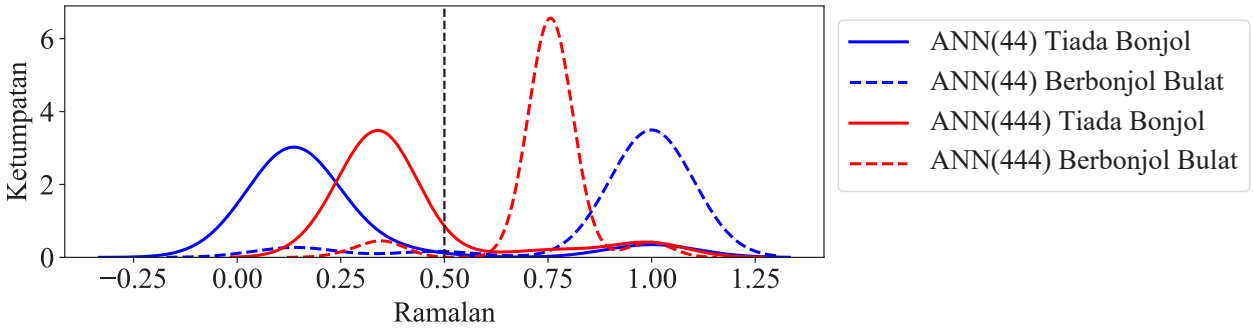
Dengan meneliti skor tertinggi bagi setiap metrik prestasi, kami mendapati bahawa diskriminan Fisher dengan transformasi penyeragaman menunjukkan prestasi paling baik, mencapai kejituan dan skor F1 tertinggi. Jurang prestasi antara Fisher dan model lain adalah masing-masing sebanyak 1.93%, 0.42% dan 3.08% untuk kejituan serta 0.62%, 0.17% dan 2.69% untuk skor F1, apabila dibandingkan dengan ANN, BDT dan *k*NN. Dari segi kepersisan, diskriminan Fisher dengan Gaussanisasi menunjukkan prestasi terbaik, mengatasi ANN, BDT dan *k*NN masing-masing sebanyak 7.26%, 4.79% dan 8.27%. Walau bagaimanapun, bagi kepekaan, ANN dengan penormalan berada di kedudukan teratas, mengatasi diskriminan Fisher sebanyak 0.62%, BDT sebanyak 0.17% dan *k*NN sebanyak 2.69%.

Berdasarkan plot KDE Gaussan dalam Rajah 4, pengelasan galaksi tanpa bonjol oleh diskriminan Fisher secara tekal menunjukkan puncak yang tinggi berbanding ANN, BDT dan *k*NN merentasi semua transformasi. Ini menunjukkan bahawa diskriminan Fisher mengelaskan galaksi tanpa bonjol dengan kejituan yang tinggi. Namun, bagi galaksi berbonjol bulat, ia menunjukkan puncak terendah merentasi semua transformasi, situasi ini tercermin pada nilai kepekaannya yang lebih rendah berbanding algoritma lain, disebabkan bilangan FN (galaksi berbonjol bulat yang salah dikelaskan sebagai tanpa berbonjol) yang tinggi. Ini bermaksud bahawa jika suatu penyelidikan yang secara khususnya ingin menapis dan membuang galaksi yang berbonjol daripada set data, ANN adalah lebih sesuai untuk menjalankan tugas ini berbanding dengan diskriminan Fisher. Pada masa yang sama, algoritma ANN dan BDT menunjukkan puncak yang lebih tinggi dalam mengelaskan galaksi berbonjol bulat berbanding diskriminan Fisher dan *k*NN. Walaupun prestasi *k*NN dalam mengelaskan galaksi berbonjol bulat adalah serupa dengan diskriminan Fisher, terdapat pertindihan ketara garis antara dua jenis galaksi, terutamanya dalam transformasi N, U dan G. Ini menunjukkan bahawa *k*NN menghadapi kesukaran dalam proses pembezaan antara dua kelas galaksi tersebut. Walaupun ANN dan BDT kurang jitu dalam pengelasan galaksi tanpa bonjol, mereka menunjukkan puncak yang hampir sama tinggi bagi kedua-dua jenis galaksi, seperti yang dapat dilihat dalam transformasi X, U dan G.

Kesimpulannya, diskriminan Fisher telah menunjukkan prestasi yang sangat baik, terutamanya dalam menangani tugas pengelasan semudah ini. Kelebihan ini mungkin berpunca daripada kerumitan model lain kerana setiap pelarasan hiperparameter algoritma boleh membawa kesan yang ketara terhadap hasilnya. Sebaliknya, struktur Fisher yang lebih ringkas membolehkannya mengekalkan prestasi yang lebih tekal. Sebagai contoh, Rajah 5 menunjukkan bahawa pemilihan benih rawak



RAJAH 4. Graf KDE Gaussan untuk jalanan tanpa transformasi (X), penormalan (N), nyahkorelasi (D), analisis komponen utama (P), penyeragaman (U) dan Gaussanisasi (G). Puncaknya menunjukkan pengelasan galaksi tanpa bonjol (0.0) dan berbonjol bulat (1.0)



RAJAH 5. Graf KDE Gaussan untuk rangkaian neural buatan (ANN) dengan benih rawak bernilai 44 dan 444 dalam jalanan tanpa transformasi (X), menunjukkan kebergantungan ANN terhadap hiperparameternya

yang berbeza akan menghasilkan lengkung KDE Gaussan berbeza: ANN dengan benih rawak bernilai 44 dan 444 menunjukkan puncak ketumpatan berada di kedudukan yang sangat berbeza dan benih rawak 44 menghasilkan ramalan yang lebih hampir kepada nilai teori dari benih rawak 444 yang lebih dekat dengan garis sempadan 0.5, meningkatkan pertindihan antara kelas. Ini menunjukkan bahawa ANN sangat sensitif terhadap nombor benih (tekal dengan penemuan Mahmud Pathi et al. 2025), yang boleh membawa kepada prestasi yang tidak tekal. Pengoptimuman nombor benih rawak ANN mengambil masa yang lama kerana beberapa kali percubaan nombor rawak perlu dilakukan untuk diambil purata prestasinya, sementara diskriminan Fisher terbukti boleh menghasilkan keputusan yang kukuh menerusi satu jalanan sahaja dalam masa yang singkat. Sebagai contoh, algoritma diskriminan Fisher kami hanya menggunakan masa 40 saat untuk mengeluarkan hasilnya, manakala ANN memerlukan 5 minit.

KESIMPULAN

Dalam kajian ini, kami telah meneroka keberkesanan diskriminan Fisher dalam menjalankan tugas pengelasan yang mudah dengan menerapkan transformasi pemboleh ubah pra-pemprosesan dan membandingkan prestasinya dengan tiga pengelas terkenal: ANN, BDT dan *k*NN. Penggunaan transformasi pemboleh ubah menunjukkan kesan yang berbeza bergantung kepada algoritma dan transformasi yang digunakan. Dari segi kejituan pengelasan antara galaksi berbonjol bulat dan tanpa bonjol, diskriminan Fisher dengan transformasi pemboleh ubah penyeragaman mencapai kejituan tertinggi (0.9310, penambahbaikan sebanyak 0.4% berbanding prestasi tertinggi algoritma lain). Bagi kepersisan, diskriminan Fisher dengan Gaussanisasi memperoleh skor tertinggi (0.9697, penambahbaikan 4.9%) manakala ANN dengan penormalan mencapai skor kepekaan tertinggi (0.9583). Secara keseluruhan, diskriminan Fisher dengan penyeragaman mencapai skor F1 tertinggi (0.9290, penambahbaikan 0.2%), menunjukkan prestasi yang cemerlang dalam pengelasan ini yang melibatkan saiz sampel yang kecil. Kami percaya bahawa ANN, BDT dan *k*NN adalah sensitif terhadap kebergantungan hiperparameter, tidak seperti diskriminan Fisher yang tidak memerlukan pelarasan hiperparameter dan lebih mudah untuk digunakan, sekali gus mengurangkan masa keseluruhan proses latihan di samping memberikan hasil yang kompetitif.

Analisis yang lebih mendalam mengenai keberkesanan diskriminan Fisher boleh dilakukan dengan meneroka julat parameter *input* yang lebih luas kerana data SDSS menyediakan banyak lagi pemboleh ubah yang masih belum diuji. Walaupun hasil kajian kami menunjukkan potensi yang ketara, adalah masih terlalu awal untuk mengangkat diskriminan Fisher sebagai penyelesaian muktamad bagi semua masalah pengelasan. Ini kerana kami percaya bahawa analisis ini mungkin bergantung pada sampel: hasil kajian kami hanya sah untuk sampel galaksi anjakan merah rendah seperti yang terdapat dalam SDSS. Walaupun penambahbaikan prestasi kejituan dan skor F1 berbanding algoritma lain adalah rendah (di bawah 1%), kami yakin bahawa diskriminan Fisher amatlah berpotensi kerana struktur algoritmanya yang ringkas serta masa jalanan yang pendek mengatasi kesemua algoritma yang telah dibandingkan dalam kajian ini. Kami juga yakin bahawa kepekaan diskriminan Fisher boleh ditingkatkan lagi jika pemilihan ciri yang lebih menyeluruh dijalankan, terutamanya penerokaan ciri galaksi yang lebih menonjolkan kehadiran bonjol suatu galaksi.

Untuk kajian masa hadapan, ujian ini boleh diluaskan kepada cabang lain dalam pokok keputusan GZ2 atau sampel data lain bagi menambah baik keputusan semasa. Penyelidikan dalam pengelasan objek samawi kekal sebagai bidang yang relevan dan penting untuk diteruskan pada masa akan datang. Kami berharap bahawa kajian ini dapat menyumbang kepada usaha dalam memahami alam semesta yang kita diami ini dengan lebih baik.

PENGHARGAAN

JYHS menyampaikan penghargaan terhadap sokongan kewangan menerusi Skim Geran Penyelidikan Fundamental (FRGS) oleh Kementerian Pengajian Tinggi Malaysia dengan kod FRGS/1/2023/STG07/USM/02/14.

RUJUKAN

Almeida, A., Anderson, S.F., Argudo-Fernández, M., Badenes, C., Barger, K., Barrera-Ballesteros, J.K. et al. 2023. The eighteenth data release of the sloan digital sky surveys: Targeting and first spectra from SDSS-V. *The Astrophysical Journal Supplement Series* 267(2): 44.

Beck, M.R., Scarlata, C., Fortson, L.F., Lintott, C.J., Simmons, B.D., Galloway, M.A., Willett, K.W., Dickinson, H., Masters, K.L., Marshall, P.J. & Wright, D. 2018. Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society* 476(4): 5516-5534.

Bertin, E. & Arnouts, S. 1996. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series* 117(2): 393-404.

Brun, R. & Rademakers, F. 1997. ROOT - An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 389(1): 81-86.

Burkert, A. & Naab, T. 2003. Major mergers and the origin of elliptical galaxies. Dlm. *Galaxies and Chaos*, *Lecture Notes in Physics 626,* disunting oleh Contopoulos, G. & Voglis, N. Springer, Berlin, Heidelberg.

Buta, R.J. 2011. Galaxy morphology. *arXiv preprint*. arXiv:1102.0550.

Chen, T. & Guestrin, C. 2016. XGBoost: A scalable tree boosting system. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. hlm. 785-794.

Dey, A., Schlegel, D.J., Lang, D., Blum, R., Burleigh, K., Fan, X., Findlay, J.R., Finkbeiner, D., Herrera, D., Juneau, S. et al. 2019. Overview of the DESI legacy imaging surveys. *Astronomical Journal* 157(5): 168.

De Diego, M.I., Redondo, A.R., Fernández, R.R., Navarro, J. & Moguerza, J.M. 2022. General performance score for classification problems. *Applied Intelligence* 52: 12049-12063.

Dickinson, H., Fortson, L., Lintott, C., Scarlata, C., Willett, K., Bamford, S., Beck, M., Cardamone, C., Galloway, M., Simmons, B., Keel, W., Kruk, S., Masters, K., Vogelsberger, M., Torrey, P. & Snyder, G.F. 2018. Galaxy zoo: Morphological classification of galaxy images from the illustris simulation. *The Astrophysical Journal* 853(2): 194.

Doi, M., Tanaka, M., Fukugita, M., Gunn, J.E., Yasuda, N., Ivezić, Ž., Brinkmann, J., de Haars, E., Kleinman, S.J., Krzesinski, J. & Leger, R.F. 2010. Photometric response functions of the sloan digital sky survey imager. *The Astronomical Journal* 139(4): 1628-1648.

Fisher, R.A. 1936. *Annals Eugenics* 7: 179.

Fraix-Burnet, D. 2023. Machine learning and galaxy morphology: For what purpose? *Monthly Notices of the Royal Astronomical Society* 523(3): 3974-3990.

Fukugita, M., Ichikawa, T., Gunn, J.E., Doi, M., Shimasaku, K. & Schneider, D.P. 1996. The sloan digital sky survey photometric system. *Astronomical Journal* 111(4): 1748.

Gunn, J.E., Siegmund, W.A., Mannery, E.J., Owen, R.E., Hull, C.L., French Leger, R., Carey, L.N., et al. 2006. The 2.5 m telescope of the sloan digital sky survey. *Astronomical Journal* 131(4): 2332-2359.

Gunn, J.E., Carr, M., Rockosi, C., Sekiguchi, M., Berry, K., Elms, B., de Haas, E. et al. 1998. The sloan digital sky survey photometric camera. *Astronomical Journal* 116(6): 3040-3081.

Hoecker, A., Speckmayer, P., Stelzer, J., Therhaag, J., von Toerne, E., Voss, H., Backes, M., Carli, T., Cohen, O., Christov, A., Dannheim, D., Danielowski, K., Henrot-Versille, S., Jachowski, M., Kraszewski, K., Krasznahorkay Jr. A., Kruk, M., Mahalalel, Y., Ospanov, R., Prudent, X., Robert, A., Schouten, D., Tegenfeldt, F., Voigt, A., Voss, K., Wolter, M. & Zemla, A. 2007. TMVA-toolkit for multivariate data analysis. *arXiv preprint* physics/0703039.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint* arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger K.Q. 2017. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. hlm. 2261-2269.

Hubble, E.P. 1926. Extragalactic Nebulae. *Astrophysical Journal* 64: 321-369.

Le, T.T., Fu, W. & Moore, J.H. 2020. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics* 36(1): 250-256.

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P. & Vandenberg, J. 2008. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society* 389(3): 1179-1189.

López-Sanjuan, C., Vázquez Ramió, H., Varela, J., Spinoso, D., Angulo, R.E., Muniesa, D., Viironen, K., Cristóbal-Hornillos, D., Cenarro, A.J., Ederoclite, A., Marín-Franch, A., Moles, M., Ascaso, B., Bonoli, S., Chies-Santos, A.L., Coelho, P.R.T., Costa-Duarte, M.V., Cortesi, A., Díaz-García, L.A., Dupke, R.A., Galbany, L., Hernández-Monteagudo, C., Logroño-García, R., Molino, A., Orsi, A., Placco, V.M., Sampedro, L., San Roman, I., Vilella-Rojo, G., Whitten, D.D., Mendes de Oliveira, C.L. & Sodré Jr., L. 2019. J-PLUS: morphological star/galaxy classification by PDF analysis. *Astronomy and Astrophysics* 622: A177.

Mahmud Pathi, I., Soo, J.Y.H., Wee, M.J., Zakaria, S.N., Ismail, N.A., Baugh, C.M., Manzoni, G., Gaztanaga, E., Castander, F.J., Eriksen, M., Carretero, J., Fernandez, E., Garcia-Bellido, J., Miquel, R., Padilla, C., Renard, P., Sanchez, E., Sevilla-Noarbe, I. & Tallada-Crespí, P. 2025. ANNz+: an enhanced photometric redshift estimation algorithm with applications on the PAU survey. *Journal of Cosmology and Astroparticle Physics* 2025: 097.

Paterson, K. 2017. MeerLICHT: MeerKAT's Optical Eye. *Proceedings of the International Astronomical Union* 14(S339): 203.

Raichoor, A., Comparat, J., Delubac, T., Kneib, J.-P., Yèche, C., Zou, H., Abdalla, F.B., Dawson, K., de la Macorra, A., Fan, X., Fan, Z., Jiang, Z., Jing, Y., Jouvel, S., Lang, D., Lesser, M., Li, C., Ma, J., Newman, J.A., Nie, J., Palanque-Delabrouille, N., Percival, W.J., Prada, F., Shen, S., Wang, J., Wu, Z., Zhang, T., Zhou, X. & Zhou, Z. 2015. The SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Selecting emission line galaxies using the Fisher discriminant. *Astronomy and Astrophysics* 585: A50.

Reza, M. 2021. Galaxy morphology classification using automated machine learning. *Astronomy and Computing* 37: 100492.

Sadeh, I., Abdalla, F.B. & Lahav, O. 2016. ANNZ2: Photometric redshift and probability distribution function estimation using machine learning. *Astronomical Society of the Pacific* 128(968): 104502.

Simmons, B.D., Lintott, C., Willett, K.W., Masters, K.L., Kartaltepe, J.S., Häußler, B., Kaviraj, S. et al. 2016. Galaxy Zoo: Quantitative Visual Morphological Classifications for 48,000 Galaxies from CANDELS. *Monthly Notices of the Royal Astronomical Society* 464(4): 4420-4447.

Stoppa, F., Bhattacharyya, S., de Austri, R., Vreeswijk, P., Caron, S., Zaharijas, G., Bloemen, S., Principe, G., Malyshev, D., Vodeb, V., Groot, P.J., Cator, E. & Nelemans, G. 2023. AutoSourceID-Classifier: Star-galaxy classification using a convolutional neural network with spatial information. *Astronomy and Astrophysics* 680: A109.

Stoughton, C., Lupton, R.H., Bernardi, M., Blanton, M.R., Burles, S., Castander, F.J., Connolly, A.J., et al. 2002. Sloan Digital Sky Survey: Early data release. *The Astronomical Journal* 123(1): 485-548.

Tan, M. & Le, Q.V. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*. PMLR 97: 6105-6114.

Urechiatu, R. & Frincu, M. 2024. Improved galaxy morphology classification with convolutional neural networks. *Universe* 10(6): 230.

Vavilova, I.B., Dobrycheva, D.V., Vasylenko, M.Y., Elyiv, A.A., Melnyk, O.V. & Khramtsov, V. 2021. Machine learning technique for morphological classification of galaxies from the SDSS: I. Photometry-based approach. *Astronomy and Astrophysics* 648: A122.

von Marttens, R., Marra, V., Quartin, M., Casarini, L., Baqui, P.O., Alvarez-Candal, A., Galindo-Guil, F.J., Fernández-Ontiveros, J.A., del Pino, A., Díaz-García, L.A., López-Sanjuan, C., Alcaniz, J., Angulo, R., Cenarro, A.J., Cristóbal-Hornillos, D., Dupke, R., Ederoclite, A., Hernández-Monteagudo, C., Marín-Franch, A., Moles, M., Sodré, L., Varela, J. & Vázquez Ramió, H. 2023. J-PLUS: Galaxy-Star-Quasar classification for DR3. *Monthly Notices of the Royal Astronomical Society* 527(2): 3347-3365.

Weglarczyk, S. 2018. Kernel density estimation and its application. *ITM web of* conferences. *EDP Sciences* 23: 00037.

Willett, K.W., Galloway, M.A., Bamford, S.P., Lintott, C.J., Masters, K.L., Scarlata, C., Simmons, B.D., Beck, M., Cardamone, C.N., Cheung, E., Edmondson, E.M., Fortson, L.F., Griffith, R.L., Haeussler, B., Han, A., Hart, R., Melvin, T., Parrish, M., Schawinski, K., Smethurst, R.J. & Smith, A.M. 2016. Galaxy Zoo: Morphological Classifications for 120,000 Galaxies in HST Legacy Imaging. *Monthly Notices of the Royal Astronomical Society* 464(4): 4176-4203.

Willett, K.W., Lintott, C.J., Bamford, S.P., Masters, K.L., Simmons, B.D., Casteels, K.R.V., Edmondson, E.M., Fortson, L.F., Kaviraj, S., Keel, W.C., Melvin, T., Nichol, R.C., Raddick, M.J. Schawinski, K., Simpson, R.J., Skibba, R.A., Smith, A.M. & Thomas, D. 2013. Galaxy zoo 2: Detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 435(4): 2835-2860.

Xie, Y., Byrd, R.H. & Nocedal, J. 2020. Analysis of the BFGS method with errors. *SIAM Journal on Optimization* 30(1): 182-209.

York, D.G., Adelman, J., Anderson Jr., J.E., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J.A. et al. 2000. The Sloan Digital Sky Survey: Technical summary. *Astronomical Journal* 120(3): 1579-1587.

Zhu, X-P., Dai, J-M., Bian, C-J., Chen, Y., Chen, S. & Hu, C. 2019. Galaxy morphology classification with deep convolutional neural networks. *Astrophysics and Space Science* 364(4): 55.

*Pengarang untuk surat-menyurat; email: johnsooyh@usm.my