Computationally-efficient Call Classification of New Zealand Birds using Texture-based Features

Yonghui Tao*, Mathis Quere†, Yusuke Hioka* and Stephen Marsland‡
* Acoustics and Vibration Research Centre, University of Auckland, Auckland, New Zealand
† SeaTech School of Engineering, University of Toulon, Toulon, France
‡ School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand
E-mail: ytao624@aucklanduni.ac.nz

***Abstract*—Bird call classification – both manual and automatic – commonly relies on spectrograms, visual time-frequency representations of the signal power distribution of the audio signal. These spectrogram "images" can be processed using deep learning approaches like convolutional neural networks (CNNs) for call identification. However, such methods typically require significant computational resources. Using labelled data from the calls of three New Zealand bird species, we compare this approach with computationally lightweight approaches that extract texture-based features from both spectrograms and audio waveforms, specifically local binary pattern (LBP) and local phase quantisation (LPQ) features, paired with classical machine learning classifiers. We find that LBP-based features outperform other feature-classifier combinations and achieve superior accuracy compared to the standard CNN approach. Given their significantly lower computational cost, texture-based techniques prove particularly well-suited for resource-constrained environments. Experimental results demonstrate that lightweight feature extraction methods can serve as efficient and effective alternatives to deep learning approaches for automated bird call classification, offering practical advantages for deployment in field conditions or embedded systems.**

I. Introduction

New Zealand's bird species face extensive population decline, with factors such as introduced predators rendering 82% of native species threatened or at risk of extinction [1]. Monitoring of species is challenging due to species being dispersed across remote, inaccessible areas and many threatened birds being cryptic in nature, creating significant obstacles for population assessment. Animal vocalisations provide effective monitoring capabilities by enabling species identification and population abundance assessment through call frequency analysis [2], [3]. To capture these vocalisations, passive acoustic recorders can be left unattended and thus provide long-time field recordings from remote areas [4], [5]. Automated methods for the analysis of these field recordings are now at a stage where they can be practically applied [6]–[8]. Convolutional neural network (CNN) architectures such as EfficientNet [9] and ResNet [10] have demonstrated reasonably high accuracy on bird call identification using large-scale datasets. Typically, the majority of these methods compute the spectrogram, a time-frequency representation of the sound based on the short-time Fourier transform (STFT), and then treat this as an image to classify calls.

Building on this foundation, there is an interest in integrating edge computing methods, which combine audio capture with on-device processing [11], [12] to provide real-time data to wildlife managers offsite. Sustaining long-term operation in harsh environments with limited power requires minimising computational costs. While CNNs are powerful they require substantial computational resources, limiting their applicability in embedded systems or remote sensing devices [13]. Further complicating their adoption is the scarcity of large, annotated datasets of New Zealand bird vocalisations required for training. These constraints motivate the exploration of alternative classification approaches that are computationally efficient and less data-dependent.

Classical machine learning (ML) methods for classification, such as k-nearest neighbours (KNN), random forest (RF), and support vector machines (SVM), are well-suited to resource-constrained audio classification problems, particularly with small datasets, as they have demonstrated strong performance in achieving highly accurate audio classification [14] while avoiding the computational overhead of deep learning approaches [15]. Unlike CNNs, these methods rely on extracting discriminative features. Among the features developed for image processing, local binary pattern (LBP) [16] and local phase quantisation (LPQ) [17] are texture-based methods that have been successfully adapted for audio, specifically environmental sound classification. Toffa and Mignotte [18] demonstrated that LBP-based classification outperformed Mel-frequency cepstral coefficients, a widely adopted feature for audio classification, in classification accuracy across KNN, RF, and SVM classifiers. Combining LBP with other audio features as input to classical ML algorithms was shown to achieve performance comparable to CNNs. The study proved the viability of computationally efficient alternatives to CNNs, while preserving competitive accuracy, which is particularly advantageous in contexts with limited computational resources and data availability.

Drawing inspiration from Toffa and Mignotte [18], this study investigates the effectiveness of using LBP/LPQ with classical ML methods, specifically for bird call classification. While the prior work [18] demonstrated the utility of LBP/LPQ with

This research was supported by funding from FlexWare Ltd, the Callaghan Innovation R&D Fellowship grant (FWRE2002/PROP-76912-FELLOW-FLEXWARE) and the University of Auckland.

979-8-3315-7206-8/25/$31.00

classical ML methods for environmental sound classification, we examine whether this approach remains viable for distinguishing between New Zealand bird species. To answer the research question, the study systematically compares the performance and computational cost of two frameworks: i) a baseline using CNNs with spectrogram inputs, and ii) a texture-based approach using classical ML methods (KNN, RF, and SVM) with LBP/LPQ features. We hypothesise that while the baseline framework may achieve superior classification accuracy, the texture-based framework will demonstrate comparable performance for bird call classification with substantially lower computational requirements, making it more suitable for resource-constrained deployment scenarios. We aim to determine the transferability of the texture-based framework to avian vocalisations, aiming to establish a potential pathway toward deployable, real-time bird call identification systems that do not rely on computationally intensive deep learning architectures.

II. Methodology

This section details the methodology including the frameworks examined and the dataset, experimental design and data analysis methods used to address the research question. Table I summarises the feature-classifer pair configurations evaluated in this study. In the texture-based framework, LBP with 1D input (LBP1), LBP with 2D input (LBP2), and LPQ, were systematically paired with KNN, RF, and SVM. The baseline framework paired spectrograms with a CNN-based architecture.

A. Texture-based Framework

The texture-based features and classical ML classifiers used in this framework are summarised. For details of each feature/classifier, readers are referred to the original literature.

a) LBP1: A 9×1 sample point neighbourhood is iteratively translated across the input waveform. The sample point values are compared with the central sample and encoded to binary values and encoded to a binary texture unit $\widetilde{S}_i$ given by:

$$\widetilde{S}_i(x) = \begin{cases} 0, & \text{if } s_x \leq s_i \\ 1, & \text{if } s_x > s_i \end{cases}, \tag{1}$$

where s denotes the sample point value, i represents the coordinates of the central sample point from the 9×1 neighbourhood on the input waveform, and $x \in \{i-4, i-3, \ldots, i+4\}$. The binary texture unit undergoes positional weighting using powers of 2, creating a weighting vector $K^T = \{2^0, 2^1, 2^2, 2^3, 0, 2^4, 2^5, 2^6, 2^7\}$. The zero weight corresponds to the central position, which is excluded from the final calculation. The decimal texture unit number is calculated by:

$$N_i = \widetilde{S}_i{}^T K, \tag{2}$$

resulting in $N_i \in \{0, 1, 2, \ldots, 255\}$.

b) LBP2: A 3×3 pixel neighbourhood is iteratively translated across the input spectrogram, which is computed by applying a 9-point Hann window with no overlap between consecutive frames. The amplitude of each pixel is compared with the central pixel and encoded to a binary texture unit given by:

$$\widetilde{G}_{ij}(x, y) = \begin{cases} 0, & \text{if } g_{x,y} \leq g_{i,j} \\ 1, & \text{if } g_{x,y} > g_{i,j} \end{cases}, \tag{3}$$

where g denotes the value of pixel, i, j represent the coordinates of the central pixel from the 3×3 window on the input spectrogram image, and $x \in \{i-1, i, i+1\}$ and $y \in \{j-1, j, j+1\}$ form a texture unit $\widetilde{G}_{ij}$. The binary texture unit undergoes positional weighting using powers of 2, creating a 3×3 weighting matrix W with elements $2^0, 2^1, 2^2, 2^3, 0, 2^4, 2^5, 2^6, 2^7$ arranged in row-major order, where the zero weight corresponds to the central pixel position. The decimal texture unit number $N_{ij} \in \{0, 1, 2, \ldots, 255\}$ is calculated as the Frobenius product $\langle \rangle_F$ of the texture unit $\widetilde{G}_{ij}$ and the weighting matrix W given by:

$$\begin{aligned} N_{ij} &= \langle \widetilde{G}_{ij}, W \rangle_F \\ &= \mathrm{Tr}(\widetilde{G}_{ij}{}^T W). \end{aligned} \tag{4}$$

c) LPQ: Using the spectrogram from LBP2, a neighbourhood of complex coefficients is formed for each spectrogram frame using the four frequency bins after the 0-th direct current bin, which correspond to frequencies up to π rad/sample:

$$F_i = \begin{bmatrix} c_1 & c_2 & c_3 & c_4 \end{bmatrix}^T, \tag{5}$$

where c_x is the spectral value in the x-th frequency bin. The spectral values are encoded to quaternary values based on their phase according to the following rule:

$$\widetilde{F}_i(x) = \begin{cases} 0, & 0 \leq \angle c_x < \frac{\pi}{2} \\ 1, & \frac{\pi}{2} \leq \angle c_x < \pi \\ 2, & \pi \leq \angle c_x < \frac{3\pi}{2} \\ 3, & \frac{3\pi}{2} \leq \angle c_x < 2\pi \end{cases}. \tag{6}$$

The quaternary values are converted into a decimal texture unit number $N_x \in \{0, 1, 2, \ldots, 255\}$ by:

$$N_x = \sum_{i=1}^{4} \widetilde{F}_i{}^T 4^{i-1}. \tag{7}$$

After applying the above operations across the entire input, the resulting texture matrix of decimal units is converted into a histogram to capture its distribution. The mean and

TABLE I
Configuration of feature-classifier pairs.

Framework	Feature	Classifier	Pairs
Texture-based	LBP1	KNN, RF, SVM	3
	LBP2	KNN, RF, SVM	3
	LPQ	KNN, RF, SVM	3
Baseline	Spec	CNN	1
Total Pairs			10

standard deviation of these histograms serve as input vectors for model training of ML methods. For the classifiers, the SVM implementation uses linear regularisation, the RF ensemble method utilises 500 decision trees, and the KNN classifier is configured with seven nearest neighbours. The code was written in Python using the Scikit-learn (v1.4.0) package.

B. Baseline Framework

To calculate the input spectrogram, the STFT was computed using a Hann window with a 128 ms frame length and no overlap between consecutive frames. The CNN model used EfficientNetV2S [19] as the base convolutional architecture, pre-trained on ImageNet [20]. The resulting feature vector was reshaped and fed into a gated recurrent unit layer with 128 hidden units, which processed the extracted features through recurrent connections to capture temporal dependencies within the feature space. The model was trained using the Adam optimiser with a learning rate of 10^{-5}. Multi-class classification was performed by a final softmax-activated dense layer. The CNN code was based on TensorFlow (v2.15.1) in Python.

C. Dataset Specifications

The experimental dataset comprised audio recordings of six distinct classes: five classes representing native New Zealand avian vocalisations and one non-call class containing natural environmental sounds without bird calls (see Table II). The avian vocal categories encompassed five call types from three endangered species: Great Spotted Kiwi or Roroa (*Apteryx maxima*), Southern Brown Kiwi or Tokoeka (*Apteryx australis*) with male and female vocalisations treated as separate classification categories, and the low-frequency boom of the Australasian Bittern or Matuku-hūrepo (*Botaurus poiciloptilus*), constituting the fifth bird call category. SBK_F samples had mixed-quality of loudness and contained much more environmental noise compared to the other call types. This dataset configuration provided a representation of endemic New Zealand birds while incorporating variability through sex-specific vocalisations, a wide distribution of call frequencies and the inclusion of a negative control class to enhance classification robustness. The audio data was derived from 15-minute continuous soundscape recordings that underwent expert annotation using the AviaNZ software package [21] to identify target vocalisations. Annotated segments corresponding to each call type were extracted to facilitate further feature computation. To standardise the temporal resolution across all data samples, call segments were cut into eight-second durations. Where necessary, audio signals were resampled to 8000 Hz to ensure a uniform sampling rate across all audio samples. For the baseline framework, spectrograms had dimensions 246×256.

D. Experimental Validation

A 20-fold cross-validation framework was implemented to ensure statistical robustness and mitigate potential overfitting. Within each fold, the dataset of six call types were partitioned into training (60%), validation (20%), and testing (20%) subsets while preserving balanced class distributions across all splits. Both training and testing process were conducted on a dedicated machine with a NVIDIA GeForce RTX 4090 graphics card, 3.19 GHz CPU, and 128 GB RAM. For computational benchmarking, inference times on the test sets were measured separately for each fold under controlled system loads on a dedicated machine. Classification performance was evaluated using F1-score, precision, and recall, derived from predictions on the held-out test sets. This protocol was systematically applied to all feature-classifier pairs, yielding computational and performance results for subsequent statistical analysis.

E. Statistical Analysis

Linear mixed-effects (LME) models for F1-score, precision, and recall were constructed to analyse the experimental results in R with the *lme4* package [22]. The LME model incorporated three categorical fixed effects: classifier, feature, and call type, with random effects accounting for repeated cross-validation measures within each combination group. Separate LME models were developed for each of the three evaluation metrics to assess performance independently. Model simplification was performed using a stepwise selection procedure to identify the most simplified yet well-fitted model using the *lmerTest* package [23]. The significance of fixed effects was evaluated via likelihood ratio tests. Two-way interactions, random intercepts and random slopes were included only when they significantly improved model fit. Following model selection, post-hoc pairwise comparisons implemented with the *emmeans* package [24] were conducted using Tukey's honest significant difference test to control for multiple comparisons. *p*-values adjusted for correction below 0.05 were considered statistically significantly.

III. Results and Discussion

A. Computational Cost

Table III presents the normalised inference time for all feature-classifier pairs. Spec_CNN is significantly slower, requiring at least 36 times more computation time compared to all other pairs. This supports the hypothesis that texture-based framework is much more lightweight than the baseline framework. Within the texture-based framework, KNN-based pairs had the lowest inference time among the classifiers and LBP1 had the lowest inference time among the features.

TABLE II
CATEGORISATION OF CALL TYPES IN THE DATASET.

Species	Sex	Class Label	Samples
Great Spotted Kiwi (Roroa, *Apteryx maxima*)	Male Female	GSK-M GSK-F	5,013 2,912
Southern Brown Kiwi (Tokoeka, *Apteryx australis*)	Male Female	SBK-M SBK-F	20,019 20,019
Australasian Bittern (Matuku-hūrepo, *Botaurus poiciloptilus*)	N/A	BITTERN	20,019
Non-call	N/A	NOISE	20,019

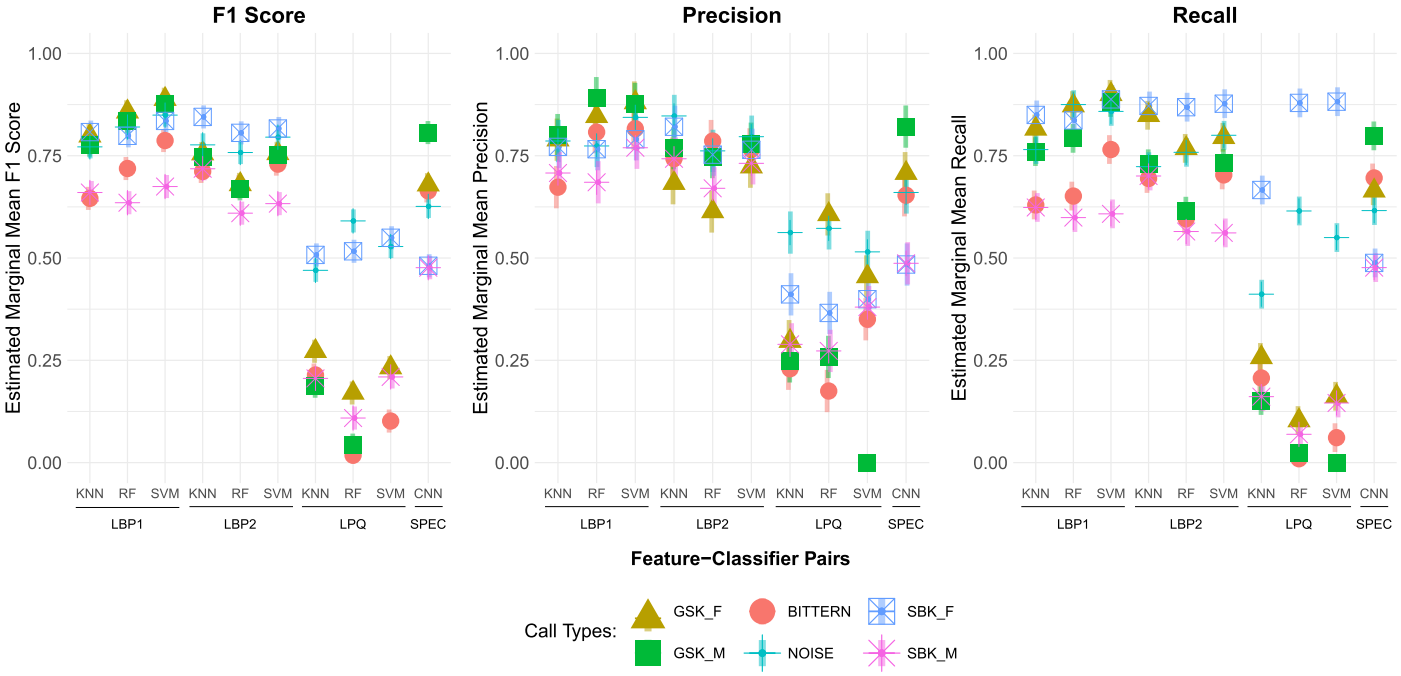

Fig. 1. Linear prediction of performance evaluation scores from LME models for feature-classifier pairs across call types (distinguished by colour and shape). Each subplot shows the corresponding scores with 95% confidence intervals, allowing comparison of classification performance across feature types and classifier algorithms for the six call types.

TABLE III
INFERENCE TIME STATISTICS COMPUTED ON A TEST SET OF 172 SAMPLES.

Pairs	Mean (s)	95% Confidence Interval (s) [lower bound, upper bound]
LBP1_KNN	0.013	[0.012, 0.014]
LBP1_RF	0.030	[0.028, 0.032]
LBP1_SVM	0.022	[0.021, 0.024]
LBP2_KNN	0.034	[0.030, 0.038]
LBP2_RF	0.049	[0.046, 0.052]
LBP2_SVM	0.047	[0.045, 0.050]
LPQ_KNN	0.015	[0.013, 0.016]
LPQ_RF	0.036	[0.034, 0.038]
LPQ_SVM	0.046	[0.044, 0.049]
Spec_CNN	1.764	[1.603, 1.925]

B. Bird Call Identification Performance

A significant two-way interaction was found between feature-classifier pairs and call types for both the F1-score ($\chi^2(45) = 1421.3, p < 0.001$) and recall ($\chi^2(45) = 1587.6, p < 0.001$). Precision had a significant two-way interaction ($F(45, 1140) = 14.774, p < 0.001$) with no random intercept. Fig. 1 shows the linear prediction of F1-score, recall, and precision. Fig. 2 shows the statistical significance of the post-hoc pairwise contrasts.

1) Texture-based framework: A consistent pattern emerges across all three metrics: LBP-based pairs demonstrate superior and comparable performance, typically achieving scores between 0.6-0.9 across most feature-classifier combinations. In contrast, LPQ-based pairs consistently shows the poorest performance across all metrics, call types, and classifiers, with scores frequently dropping below 0.5 and in some cases approaching zero, particularly evident in the GSK-M and SBK-M where it fails almost entirely across all metrics.

Across all metrics, LBP1_SVM emerged as the best performing combination for most call types, followed by LBP2_KNN and LBP1_RF. For F1-score, LBP1_SVM achieved the highest F1-score for GSK_M/F, BITTERN and NOISE call types. LBP2_KNN had the highest F1-score for SBK_M while SBK_F had very few statistically significantly differences between the top performing LBP-based pairs. Comparing LBP1 with LBP2 pairs, there were few to no significant differences in F1-score for SBK_M/F, BITTERN and NOISE. While, LBP1 pairs scored statistically significantly higher for GSK_M/F. For precision, the top performing pairs were LBP1_SVM (BITTERN, GSK_F, SBK_M), LBP1_RF (GSK_M) and LBP2_KNN (NOISE, SBK_F). Most call types had few or no significant differences apart from GSK_F where LBP1 had statistically significantly higher scores than LBP2 pairs. For recall, the top performing pairs were LBP1_SVM (BITTERN, GSK_F, GSK_M), LBP1_RF (NOISE) and LBP2_KNN (SBK_F, SBK_M). Only GSK_M and NOISE showed some LBP1 pairs had statistically significantly higher scores, while the remaining calls had few to no significant differences.

2) Comparison between frameworks: The baseline framework shows moderate performance, generally ranging from 0.5-0.8. Comparing Spec_CNN with LBP-based pairs, Spec_CNN performed worse than most LBP-based pairs with statistically significantly lower scores in F1-score for SBK_M/F, NOISE and GSK_F, while GSK_M and BITTERN showed less significant differences. For precision, SBK_M/F

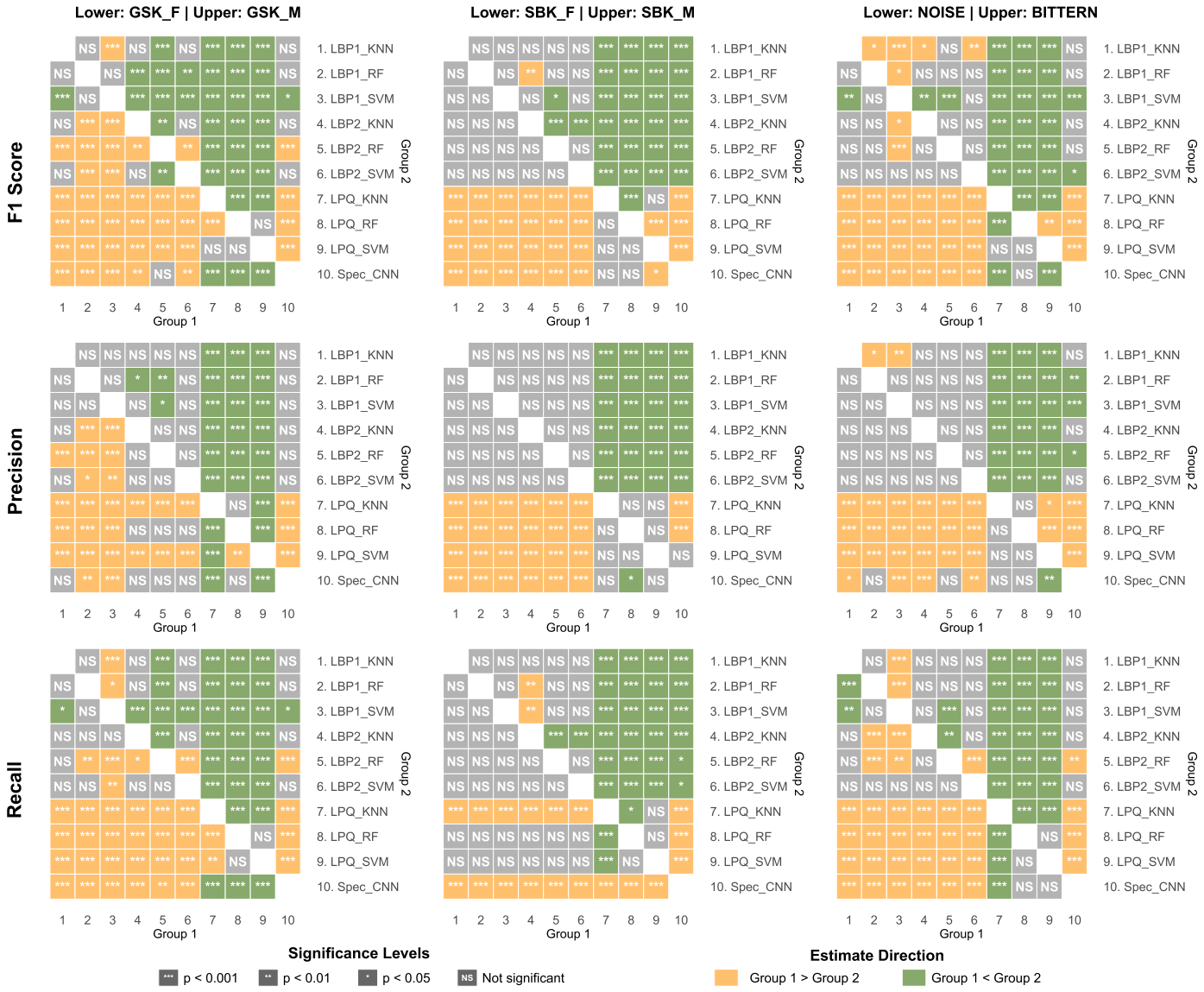

Fig. 2. Combined post-hoc pairwise contrasts displayed according to evaluation metric (row), call type (column separated into upper and lower triangles) and feature-classifier pairs (shown in group 1/2). The *p*-values are represented by significance levels and the estimate direction represents whether group 1 performed better (>) or worse (<) than group 2.

and GSK_M had statistically significantly lower scores than LBP-based pairs. For recall, Spec_CNN had statistically significantly lower scores than LBP-based pairs for most call types, except BITTERN and GSK_M. Comparing Spec_CNN with LPQ-based pairs, Spec_CNN had statistically significantly higher scores than LPQ-based pairs for most call types apart from SBK_F and NOISE where some pairs had no significant difference. For precision, Spec_CNN had statistically significantly higher scores than all LPQ-based pairs for BITTERN and GSK_M. For recall, most call types had had statistically significantly higher scores than LPQ-based pairs, while for SBK_F, Spec_CNN had statistically significantly lower scores than LPQ-based pairs.

C. Discussion

Our hypothesis that the texture-based framework would achieve comparable performance to the baseline framework while offering reduced computational complexity was partially supported by the experimental results. Regarding computational efficiency, the texture-based framework resulted in substantially reduced inference times compared to the baseline framework, confirming the hypothesis. This result was expected due to the lightweight classical ML classifiers and the lower-dimensionality texture-based features. For classification performance, LBP-based pairs demonstrated competitive results, either matching or exceeded those of Spec_CNN, establishing LBP-based approaches as the more advantageous method for the dataset and scale examined in this study.

The study by Toffa and Mignotte [18] showed that LBP2 outperformed both LBP1 and LPQ, while LBP1 and LPQ exhibited comparable performance levels. In contrast, our results revealed that LBP1 and LBP2 performed similarly, while LPQ showed markedly inferior performance across most metrics. This discrepancy indicates that the performance of texture-based features is dependent on the dataset, likely due to differences in signal characteristics between environmental sounds and bird vocalisations, particularly in their spectral and temporal properties. The superior performance of LBP2 can be attributed to its more complete time-frequency representation of the signal, which enhances robustness. Notably, LBP1 achieved comparable results despite relying solely on 1D signal variation. While LPQ is more robust to noise, the consistent low recall across most call types suggests that focus on noise-handling may actually hinder its ability to capture subtle patterns in cleaner recordings. Future research will explore the effect of feature collaboration with classical ML classifiers on the performance of bird call classification.

IV. Conclusion

This study evaluated the transferrability of using texture-based features with classical ML methods for distinguishing between New Zealand bird species. The study compared the computational and classification performance of texture-based and baseline frameworks. The experimental results confirm that texture-based framework is a viable, computationally-efficient alternative to the baseline for bird call classification. Specifically, LBP-based feature-classifier pairs, notably LBP1_SVM, achieved comparable or better performance than the baseline with significantly reduced computational complexity. The ability to achieve comparable classification accuracy with substantially lower computational overhead makes these methods particularly suitable for monitoring applications in remote conservation contexts.

Acknowledgement

We are grateful to Dr. Justine Hui for guidance on the statistical analyses presented in this work.

References

[1] Stats NZ. "Extinction threat to indigenous species." (2021).

[2] J. Wimmer, M. Towsey, P. Roe, and I. Williamson, "Sampling environmental acoustic recordings to determine bird species richness," *Ecological Applications*, vol. 23, no. 6, 2013.

[3] J. P. G. Jones, "Monitoring species abundance and distribution at the landscape scale," *Journal of Applied Ecology*, vol. 48, no. 1, 2011.

[4] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, and D. Llusia, "Terrestrial Passive Acoustic Monitoring: Review and Perspectives," *BioScience*, vol. 69, no. 1, Jan. 1, 2019.

[5] E. M. Williams, C. F. J. O'Donnell, and D. P. Armstrong, "Cost-benefit analysis of acoustic recorders as a solution to sampling challenges experienced monitoring cryptic species," *Ecology and Evolution*, vol. 8, no. 13, 2018.

[6] S. Kahl, T. Denton, H. Klinck, *et al.*, "Overview of BirdCLEF 2024: Acoustic Identification of Under-studied Bird Species in the Western Ghats,"

[7] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, 2019.

[8] A. Maithripala, S. M. Arachchi, K. Karunanayaka, R. Perera, and P. Pallewatta, "A Review of Automated Bird Sound Recognition and Analysis in the New AI Era," in *2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, Dec. 2024.

[9] M. U. Sheikh, H. Abid, B. S. Shafique, A. Hanif, and M. H. Khan, "Bird Whisperer: Leveraging Large Pre-trained Acoustic Model for Bird Call Classification," in *Interspeech 2024*, ISCA, Sep. 1, 2024.

[10] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, Mar. 1, 2021.

[11] O. Kücüktopcu, E. Masazade, C. Ünsalan, and P. K. Varshney, "A real-time bird sound recognition system using a low-cost microcontroller," *Applied Acoustics*, vol. 148, May 1, 2019.

[12] Z. Huang, A. Tousnakhoff, P. Kozyr, *et al.*, "TinyChirp: Bird Song Recognition Using TinyML Models on Low-power Wireless Acoustic Sensors," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Sep. 2024.

[13] X. Luo, D. Liu, H. Kong, *et al.*, "Efficient Deep Learning Infrastructures for Embedded Computing Systems: A Comprehensive Survey and Future Envision," *ACM Trans. Embed. Comput. Syst.*, vol. 24, no. 1, Dec. 10, 2024.

[14] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, 4 Sep. 29, 2020.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York: Springer, 2006.

[16] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, Oct. 1994, 582–585 vol.1.

[17] V. Ojansivu and J. Heikkilä, "Blur Insensitive Texture Classification Using Local Phase Quantization," in *Image and Signal Processing*, Springer, 2008, pp. 236–243.

[18] O. K. Toffa and M. Mignotte, "Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3978–3985, 2021.

[19] M. Tan and Q. V. Le. "EfficientNetV2: Smaller Models and Faster Training." arXiv: 2104.00298 [cs]. (Jun. 23, 2021), pre-published.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[21] S. Marsland, N. Priyadarshani, J. Juodakis, and I. Castro, "AviaNZ: A future-proofed program for annotation and recognition of animal sounds in long-time field recordings," *Methods in Ecology and Evolution*, vol. 10, no. 8, 2019.

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, pp. 1–48, 2015.

[23] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "Lmertest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, pp. 1–26, 2017.

[24] R. V. Lenth, P. Buerkner, M. Herve, *et al.*, *Emmeans: Estimated marginal means, aka least-squares means*, R package, 2022.