Seminar Nasional Instrumentasi, Kontrol dan Otomasi (SNIKO) 2018
Bandung, Indonesia, 10-11 Desember 2018

Evaluasi Performa metode *Deep Learning* untuk Klasifikasi Citra Lesi Kulit *The HAM10000*

[1]Harits Abdurrohman*), [1]Robih Dini & [2]Arief Purnama Muharram

[1]Fakultas Ilmu Komputer Universitas Brawijaya

[2]Fakultas Kedokteran Universitas Indonesia

(cooresponding author) harits.abdr@gmail.com)*

Abstrak

The HAM10000 Dataset merupakan koleksi besar citra dermatoskopi untuk lesi kulit berpigmen yang umum. *The HAM10000 Dataset* terdiri atas 10.015 data citra lesi kulit berpigmen yang terbagi untuk penyakit Bowen, karsinoma sel basal, *benign keratosis-like lesion*, dermatofibroma, melanoma, *melanocytic nevi*, dan lesi vaskular. Data citra yang terdapat dalam dataset telah terkonfirmasi baik melalui histopatologi, pemeriksaan *follow-up*, konsensus pakar, maupun konfirmasi melalui *in-vivo confocal microscopy*. Pada penelitian ini kami melakukan pengujian performa terhadap model *deep learning* dan melakukan evaluasi. Tahap *pre-processing* citra meliputi analisis distribusi citra pada setiap kelas lesi, pengelompokan ulang kelas lesi berdasarkan letak pada bagian tubuh, dan augmentasi citra. Oleh karena keterbatasan data citra setelah dilakukan analisis distribusi maka model yang dibangun pada penelitian ini hanya berfokus pada kelas lesi untuk abdomen, punggung, ekstremitas atas dan bawah. Evaluasi ini dilakukan terhadap beberapa metode yang terkenal *InceptionV3*, *MobileNet* dan *MobileNetV2.* Ukuran performa yang dilakukan meliputi analisis *confusion matrix* yakni dengan mengambil nilai *precision dan recall*, dan *f1-score*.

Kata Kunci: Deep Learning, Performance, The HAM10000 Dataset, Skin Lesion

1 Pendahuluan

1.1 Latar Belakang

Dermatoskopi digunakan untuk membantuk melakukan diagnosa terhadap lesi kulit berpigmen sebagai pembanding dari pemeriksaan dari mata telanjang [8]. Untuk melakukan diagnosa terhadap lesi kulit, ada beberapa cara yang dapat dilakukan yakni dengan histopatologik. Histopatologik merupakan teknik pemeriksaan mikroskopik terhadap suatu jaringan untuk melihat perkembangan dari suatu penyakit dimana jaringan tersebut sudah dipindahkan dalam suatu wadah yang diberi pengawet untuk mencegah pembusukan. Selain histopatologik, metode lain yang dapat dilakukan adalah dengan *in-vivo confical microscopy.* Teknik ini merupakan teknik pencitraan dengan resolusi mendekati tingkat sel yang pada umumnya *facial benign keratoses* terbantu dengna metode ini [11]. Lesi kulit terkadang membutuhkan waktu 1.5 hingga 3 tahun untuk mengetahui dan membuktikan bahwa lesi kulit tersebut jinak. Pemeriksaan ini disebut dengan pemeriksaan *follow-up.* Sedangkan cara lain adalah melakukan konsensus kepada para ahli. Lesi dengan jenis *ground-truth* ini biasanya difoto untuk alasan pendidikan dan tidak perlu tindak lanjut atau biopsi untuk konfirmasi[13].

The HAM10000 (*Human Against Machine*) dataset merupakan dataset yang berisi citra dermatoskopik untuk lesi kulit berpigmen umum. Kumpulan citra ini dikumpulkan selama 20 tahun dari dua tempat berbeda, yakni dari Departemen Dermatologi di *Medical university of Vienna,* Austria dan praktik kanker kulit Cliff Rosendahl di Queensland, Australia. Dataset ini terdiri dari 10.015 citra dermatoskopik [13]. Tujuan dari dibentuknya dataset ini sendiri sebagai data latih bagi machine learning untuk kebutuhan akademik. Data yang diperoleh ini telah divalidasi oleh para ahli dengan empat jenis validasi yakni histopatologi, konfokal, konsensus pakar dan pemeriksaan *follow-up.* Dengan jumlah data yang banyak, *the HAM10000* dapat dijadikan sebagai data latih untuk metode *machine learning.*

Machine learning terdiri dari beberapa sub bagian, yakni *supervised, unsupervised* dan *reinforcement learning.* Masing-masing bagian memiliki pendekatannya sendiri. Algoritme *supervised-learning* terbagi lagi menjadi beberapa sub berdasarkan pendekatannya, yaitu *logic-based algorithms, perceptron-based algorithms,* dan *statistical learning algorithms* [6]. Dalam pengembangannya, metode dari *machine learning* tersebut berubah sesuai dengan kebutuhan atau mengikuti dari bentuk data yang tersedia, seperti *sequential*, *time series*, berupa teks maupun foto. Pada tahun 1986, istilah *deep learning* muncul sebagai salah satu dari metode *machine learning* oleh Rina Dechter [2] kemudian beralih ke sub bagian lain terkhusus *perceptron-based algorithms,* yakni jaringan syaraf tiruan.

Pada tahun 1998, Yann LeCun merancang sebuah arsitektur yang disebut dengan *convolutional neural network* yang bekerja dengan data berupa citra [7]. *Convolution neural network* sendiri terdiri

dari layer konvolusi yang mengekstraksi tekstur dari sebuah citra. Diikuti dengan *backpropagation* untuk memperbarui bobotnya.

Seiring berjalannya waktu, arsitektur dari *deep learning* semakin beragam. Salah satunya adalah inception V3 yang merupakan arsitektur pertama dengan parameter yang lebih sedikit dan komputasi yang efisien [12]. Pada inception V3 ini ada faktorisasi untuk mengurangi parameter. Meskipun komputasinya sudah rendah, inception V3 tidak dapat digunakan pada perangkat dengan ruang lingkup komputasi rendah. Pada tahun 2017, Mobilenet diciptakan oleh Google untuk menjawab masalah tersebut. Mobilenet mempunyai layer khusus yang disebut dengan *depthwise separable convolution.* Layer *depthwise separable convolution* ini digunakan untuk mereduksi kompleksitas dan lebih sedikit parameter sehingga menghasilkan model yang lebih ukurannya. Perkembangan lebih lanjut dari mobilenet adalah mobilenet V2. Pada mobilenet V2, dua fitur terbarunya adalah *linear bottleneck* pada setiap layer dan koneksi *shortcut* antara *bottlenecks* [9].

1.2 Tujuan

Merujuk pada penelitian Binder pada tahun 1994 [1], bahwa citra dermatoskopi dapat digunakan sebagai data latih untuk jaringan syaraf tiruan, maka pada penelitian ini kami melakukan eksperimen yang serupa namun dengan model yang berbeda. Tujuan dari penelitian ini adalah menguji ketiga model *deep learning* (mobilenet V1, mobilenet V2, dan inception V3) dalam melakukan klasifikasi jenis lesi kulit berpigmen umum. Adapun pengukuran yang dilakukan dengan membandingkan nilai *precision, recall* dan *F-1 score* serta analisis dari *confusion matrix.*

2 Landasan Kepustakaan dan Metodologi Penelitian

2.1 Factorizing Convolutions

Factorizing convolutions merupakan *novelty* yang dihadirkan pada model deep learning inception. Factorizing convolutions melakukan faktorisasi terhadap kernel konvolusi. Tujuan dari faktorisasi ini untuk mereduksi jumlah dari parameter yang dihasilkan dari setiap layer [12]. Faktorisasi ini bisa dilakukan dengan dua cara yakni:

1. Faktorisasi menjadi konvolusi yang lebih kecil.
   Proses yang dilakukan adalah mengganti ukuran kernel konvolusi menjadi ukuran yang lebih kecil. Sebagai contoh jika menggunakan filter berukuran 5×5 akan menghasilkan 25 parameter. Filter ini bisa digantikan dengan menggunakan 2 kernel berukuran 3×3 dan menghasilkan 18 parameter saja, dimana hal tersebut mereduksi jumlah parameter sebanyak 28%.
2. Faktorisasi menjadi konvolusi yang asimetrik.
   Berbeda dengan faktorisasi sebelumnya, faktorisasi ini membagi ukuran kernel dengan faktornya. Sebagai contoh filter dengan ukuran kernel 3×3 difaktorisasi menjadi 3×1 dan 1×3. Ketika filter dengan kernel berukuran 3×3 menghasilkan 9 parameter, dua filter hasil faktorisasinya yakni $3 \times 1 + 1 \times 3$ menghasilkan 6 parameter, yang berarti jumlah parameter tersebut tereduksi sebanyak 33%.

2.2 Depthwise Separable Convolution

Depthwise separable convolution merupakan *novelty* yang dihadirkan pada model mobilnet V1. *Depthwise separable convolution* merupakan sebuah blok pada *deep learning* yang terdiri dari *depthwise convolution* dan *pointwise convolution.* Tujuan dari *depthwise separable convolution* ini untuk mereduksi komputasi dan ukuran dari model [4]. *Depthwise separable convolution* sendiri diciptakan pada tahun 2014 sebagai disertasi [10]. *Depthwise convolution* merupakan hasil faktorisasi dari konvolusi standar. Dari N jumlah input, *depthwise convolution* melakukan prosesnya untuk setiap kanalnya. Sebagai contoh, input dari layer *depthwise convolution* ada 10 kanal, maka akan menghasilkan 10 hasil konvolusi baru. *Pointwise convolution* merupakan kernel dengan ukuran 1×1 yang digunakan untuk menggabungkan seluruh hasil konvolusi dari *depthwise convolution.* Berikut adalah total biaya operasi yang dilakukan, dimana bagian kiri adalah *depthwise convolution* dan bagian kanan adalah *pointwise convolution:*

$$D_k \cdot D_k \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \qquad (1)$$

Dimana D_k adalah ukuran atau dimensi dari kernel, M adalah jumlah kanal input, N adalah jumlah dari kanal output dan D_F adalah ukuran dari fitur atau filter. Sedangkan untuk konvolusi standar, biaya komputasinya adalah sebagai berikut:

$$D_k \cdot D_k \cdot M \cdot N \cdot D_F \cdot D_F \qquad (2)$$

Pada konvolusi standar, operasi konvolusi dilakukan oleh setiap filter dengan seluruh kanal. Dengan melakukan faktorisasi ini, *depthwise separable convolution* mereduksi biaya komputasi

tersebut. Berikut adalah reduksi dari komputasi yang dilakukan oleh *depthwise separable convolution:*

$$\frac{D_k \cdot D_k \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_k \cdot D_k \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_k^2} \quad (3)$$

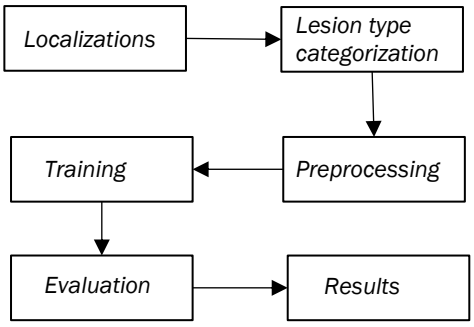
Maka dapat dikatakan jika ukuran dari kernel $D_k \times D_k$ adalah 3×3, maka komputasinya akan tereduksi 8 hingga 9 kali lebih sedikit.

2.3 Linear Bottlenecks

Linear bottlenecks adalah sebuah blok yang diciptakan pada model mobilenet V2. *Linear bottleneck* menggunakan *depthwise separable convolution* untuk membantu dalam mereduksi komputasi. *Linear bottleneck* sendiri dibentuk dari blok yang berkebalikan dari *residual block,* yakni *inverted residual block*. *Residual block* membantu untuk mengembalikan informasi yang hilang saat aktivasi oleh ReLU dengan menggunakan *skip connection[*3]. *Linear bottleneck* menghapus aktivasi ReLU di akhir proses *inverted residual block*. Sehingga informasi yang diberikan tidak ada yang hilang.

2.4 Metodologi Penelitian

Metodologi penelitian yang kami kerjakan antara lain adalah sebagai berikut:



Gambar 1 Diagram Alir Metodologi Penelitian

Berdasarkan Gambar 1, dalam mengolah dataset *The HAM10000* yang pertama kali dilakukan adalah melakukan lokalisasi, yakni membagi data sesuai dengan lokasi tubuh yang ada. Adapun dari beberapa bagian yang diberikan, pada penelitian ini kami memilih empat lokasi tubuh, yaitu perut (abdomen), ekstremitas atas, ekstremitas bawah dan punggung. Kedua, setelah melakukan lokalisasi, pada masing-masing lokasi tersebut kami membaginya berdasarkan kategori tipe lesinya. Adapun tipe lesi kulit yang digunakan adalah penyakit Bowen, karsinoma sel basal, *benign keratosis-like lesion*, melanoma, dermatofibroma, *melanocytic nevi*, dan lesi vaskular. Ketiga, kami melakukan *preprocessing* terhadap data yang telah dibagi tersebut karena persebarannya yang tidak merata pada setiap kelas. *Preprocessing* yang dilakukan adalah melakukan augmentasi data. Augmentasi data adalah proses penggandaan data dengan melakukan translasi, transformasi, penambahan *noise,* rotasi, pembesaran, atau *flipping* [5]. Augmentasi data membantu untuk menambah data jika datanya terlalu sedikit untuk kebutuhan *deep learning*. Masing-masing citra juga diubah ukurannya menjadi 224×224 piksel. Keempat melakukan proses pelatihan terhadap model dari *deep learning*. Adapun proses yang dilakukan disamakan untuk semua model yang diujikan. Proses pelatihan dilakukan hingga nilai dari *learning rate* yang dihasilkan telah mencapai 0.00001. Proses pelatihan dilakukan di masing-masing lokasi yang telah dibagi. Kelima adalah evaluasi dari masing-masing model untuk setiap lokasi lesi kulit. Pada bagian ini kami mencoba melakukan analisis terhadap *confusion matrix* yang dihasilkan serta satuan metrik pengujian yakni *precision, recall,* dan *F-1 score*. Terakhir, kami menyajikan hasilnya dalam pembahasan.

3 Hasil dan Pembahasan

Pada bagian ini kami menyampaikan hasil dari evaluasi dalam penelitian ini.

3.1 Performa

Hasil pengujian terhadap performa dari ketiga model ditampilkan pada Tabel 1 sampai dengan Tabel 4. Ketiga metrik pengujian ini memiliki tujuannya masing-masing. *Recall* menunjukkan jumlah objek dalam suatu kelas secara aktual dan prediksi adalah benar atau sama sedangkan *precision* menunjukkan jumlah objek yang dipilih adalah benar. *F-1 score* menunjukkan rata-rata harmonik dari *recall* dan *precision*. Dalam kehidupan nyata jika nilai *recall* terlalu rendah maka akan sangat berbahaya jika diaplikasikan karena *recall* menentukan suatu objek dikenali sebagai objek yang berbahaya atau tidak. *Precision* akan baik digunakan jika model dapat membedakan antara aktual negatif dengan aktual positif. Jika aktual negatif diprediksi banyak diprediksi dengan benar oleh model, maka kemungkinan suatu objek yang tidak berbahaya dapat dikenali sebagai objek yang berbahaya. *F-1 score* membantu mengetahui kemampuan model dalam mengenali *false negative* dan *false positive*.

Tabel 1 Nilai rata-rata pengujian pada abdomen

	Precision	*Recall*	*F1-score*
Mobilenet	0.76	0.32	0.37
Mobilenet V2	0.69	0.81	0.75
Inception V3	0.77	0.80	0.76

Tabel 2 Nilai rata-rata pengujian pada punggung

	Precision	*Recall*	*F1-score*
Mobilenet	0.71	0.69	0.65
Mobilenet V2	0.66	0.62	0.60
Inception V3	0.55	0.65	0.58

Tabel 3 Nilai rata-rata pengujian pada ekstremitas bawah

	Precision	*Recall*	*F1-score*
Mobilenet	0.85	0.71	0.73
Mobilenet V2	0.62	0.69	0.62
Inception V3	0.65	0.59	0.61

Tabel 4 Nilai rata-rata pengujian pada ekstremitas atas

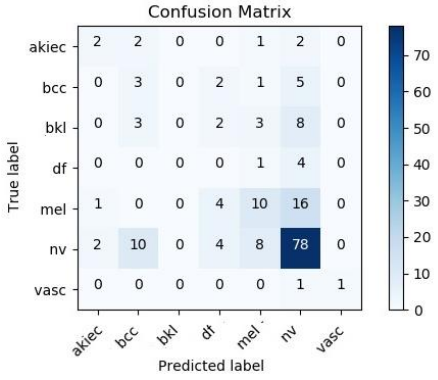
	Precision	*Recall*	*F1-score*
Mobilenet	0.66	0.66	0.64
Mobilenet V2	0.51	0.54	0.52
Inception V3	0.55	0.51	0.50

Dari hasil pengujian yang dilakukan, nilai *precision* terendah ada pada lokasi ekstremitas atas dengan nilai hanya 51% dapat mengenali tipe lesi kulit pada model Mobilenet V2. Nilai *recall* paling rendah dihasilkan oleh model mobilenet pada lokasi abdomen dengan nilai 32%. Ini menunjukkan model buruk dalam mengenali tipe lesi kulit. Nilai *F-1 score* terendah ada pada model mobilenet pada lokasi abdomen. Nilai ini menujukkan bahwa model ini memiliki nilai *false negative* dan *false positive* yang besar.

Sedangkan untuk nilai terbaik pada *precision* ada pada model mobilenet di lokasi ekstremitas bawah dan *recall* terbaik ada pada model mobilenet V2 di lokasi abdomen. Untuk nilai *F-1 score* terbaik ada pada model inception V3 pada lokasi abdomen. Nilai-nilai ini dapat memberikan gambaran bahwa model yang dibangun belumlah baik. Oleh karena itu dari nilai-nilai tersebut, kami melakukan analisis dari *confusion matrix* yang dihasilkan.

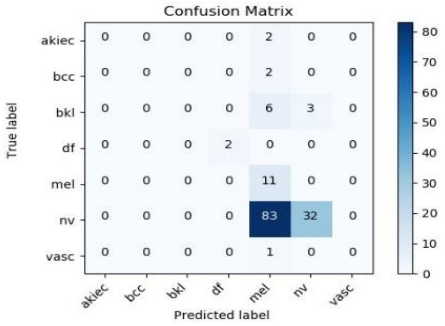
3.2 Analisis

Analisis dilakukan dengan merujuk pada *confusion matrix yang* dihasilkan oleh masing-masing model. Pada penelitian kami memfokuskan pada nilai terburuk dan terbaik dari pengujian performa yang telah dilakukan.
pre



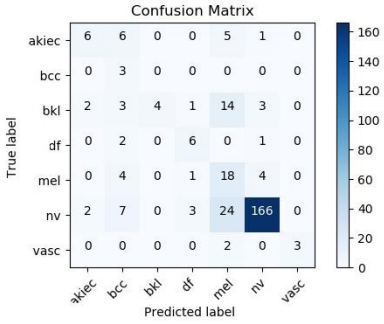
Gambar 2 *Confusion martrix* ekstremitas atas mobilenet V2

Pada Gambar 2, dapat dilihat bahwa terdapat kecenderungan model dapat mengenali lesi kulit nevi dan untuk beberapa tipe lesi lainnya dikenali salah, yakni pada dermatofibroma, *benign keratosis-like lesion,* dan melanoma. Pada model ini, melanoma hanya dapat dikenali sebanyak 62.5% sedangkan jenis tipe lesi dermatofibroma dan *benign keratosis-like lesion* bernilai 0%. Model ini tidak dapat mengenali citra lesi tersebut dengan baik dikarenakan oleh jumlah data yang terlalu sedikit atau ada kemiripan dari tesktur dengan kelas yang dikenalinya.
Abdomen V1 recall



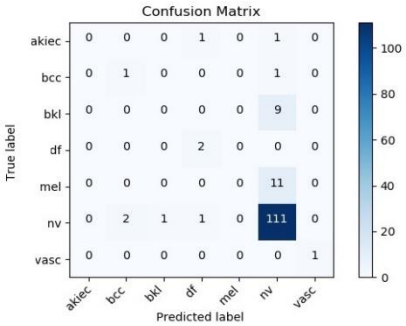
Gambar 3 *Confusion matrix* abdomen mobilenet

Pada Gambar 3 dapat disimpulkan bahwa model mobilnet untuk abdomen sangatlah buruk dengan hanya ada tiga kelas saja yang dapat dikenali dengan baik. Citra lesi nevi banyak diklasifikasikan sebagai melanoma, yang berarti model ini mempunya false negative error yang tinggi. Model ini tidak baik digunakan pada kasus nyata.



Gambar 4 *Confusion matrix* ekstremitas bawah mobilenet

Gambar 4 merupakan contoh dari model yang baik karena semua kelas dapat dikenali dengan baik dengan nilai galat yang lebih sedikit dibandingkan dengan model pada Gambar 2 dan Gambar 3. Model tersebut dapat mengenali karsinoma sel basal dengan baik. Nilai galat paling besar ada pada kelas benign keratosis-like lesion, dengan nilai false negative error sebesar 85.18%. Dari model ini juga dapat diambil kesimpulan bahwa citra lesi melanoma berbagi nilai fitur yang sama terhadap citra lesi lainnya yang menyebabkan model memiliki kecenderungan mengidentifikasi sebuah citra lesi dengan melanoma.



Gambar 5 *Confusion matrix* abdomen mobilenet V2

Gambar 5 memiliki hasil yang sama dengan Gambar 3, hanya saja nilai recall yang dihasilkan lebih tinggi. Hal ini berkat model dapat mengenali tipe lesi nevi dengan baik. Namun hal tersebut tidak diikuti dengan *precision* yang baik. Dengan hanya 4 kelas yang dapat dikenali, model ini tidak layak digunakan pada kehidupan nyata, karena akan berbahaya jika suatu lesi tidak dikenali sama sekali oleh model tersebut.

4 Kesimpulan dan Saran

Kesimpulan yang diperoleh dari penelitian ini adalah model yang baik adalah model dengan nilai *F-1 score* terbaik, karena hal tersebut menentukan nilai dari *precision* dan *recall* dari model yang bersangkutan. Namun pada realitanya, nilai tersebut perlu dibuktikan dengan *confusion matrix.* Dari penelitian kami menyimpulkan model terbaik ada pada mobilenet untuk lokasi ekstremitas bawah, ekstremitas atas, dan punggung. Kemudian model mobilenet V2 untuk lokasi abdomen. Masing-masing lokasi tidak dapat diambil model yang terbaik karena jumlah data yang digunakan juga berbeda. Ketiga model baik mobilenet, mobilenet V2 dan inception V3 masing-masing memiliki keunggulannya sendiri, tapi karena terbatas oleh dataset yang rendah dengan sedikit variasi menyebabkan nilai yang diperoleh jauh dari ekspektasi sebelumnya. Hasil dari model penelitian tidak dapat digunakan untuk kasus nyata terkait masih banyaknya hal yang perlu dianalisis lebih lanjut.

Adapun saran untuk penelitian selanjutnya adalah menggunakan dataset *The HAM10000* dengan langsung membaginya berdasarkan kategorinya saja karena informasi seperti testur dari kulit tidak dapat diekstraksi dengan model *deep learning* jika input atau ukuran dari citra tersebut berukuran 224×224 piksel. Saran lain berupa membuat model yang dibangun atas layer-layer yang ada pada ketiga model yang diujikan pada penelitian in.

5 Ucapan Terima kasih

Kami mengucapkan terima kasih kepada rekan kami Ibrahim Aji, Nanda Permata Putri, dan Febrian Falentino Fredriktho pada pelatihan Digitalent 2018 yang diadakan oleh Kemkominfo karena telah membantu dalam melakukan penelitian ini.

6 Daftar Pustaka

[1] Binder, M. et al., 1994. Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. British Journal of Dermatology, 4(130), pp. 460-465.

[2] Dechter, R., 1986. Learning While Searching in Constraint-Satisfaction-Problems. University of California, Computer Science Department, Cognitive Systems Laboratory.

[3] He, K., Zhang, X., Ren, S. & Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition , pp. 770-778.

[4] Howard, A. G. et al., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint .

[5] Hussain, Z., Gimenez, F., Yi, D. & Rubin, D., 2017. Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. AMIA Annual Symposium Proceedings, Volume 2017, p. 979.

[6] Kotsiantis, S. B., Zaharakis, I. & Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, Volume 160, pp. 3-24.

[7] LeCun, Y., Haffner, P., Bottou, L. & Bengio, Y., 1999. Object recognition with gradient-based learning. Shape, contour and grouping in computer vision, pp. 319-345.

[8] Rosendahl, C., Tschandl, P., Cameron, A. & Kittler, H., 2011. Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. J Am Acad Dermatol, Volume 64, p. 1068–1073.

[9] Sandler, M. et al., 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510-4520.

[10] Sifre, L. & Mallat, S., 2014. Rigid-motion scattering for image classification. Doctoral dissertation, PhD thesis, Ph. D. thesis.

[11] Stevenson, A. D., Mickan, S., Mallett, S. & Ayya, M., 2013. Systematic review of diagnostic accuracy of reflectance confocal microscopy. Dermatol Pract Concept, Volume 3, p. 19–27.

[12] Szegedy, C. et al., 2016. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826.

[13] Tschandl, P., Rosendahl, C. & Kittler, H., 2018. The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. arXiv preprint .